SEMI-PARAMETRIC ESTIMATION OF GENERALIZED PARTIALLY LINEAR SINGLE-INDEX MODELS

Yingcun Xia
Department of Zoology, University of Cambridge, UK

Wolfgang Härdle
CASE-Center for Applied Statistics and Economics
Humboldt-Universität zu Berlin

One of the most difficult problems in applications of semiparametric generalized partially linear single-index model (GPLSIM) is the choice of pilot estimators and complexity parameters which may result in radically different estimators. Pilot estimators are often assumed to be root-n consistent, although they are not given in a constructible way. Complexity parameters, such as a smoothing bandwidth are constrained to a certain speed, which is rarely determinable in practical situations.

In this paper, efficient, constructible and practicable estimators of GPLSIMs are designed with applications to time series. The proposed technique answers two questions from Carroll *et al.* (1997): no root-n pilot estimator for the single index part of the model is needed and complexity parameters can be selected at the optimal smoothing rate. The asymptotic distribution is derived and the corresponding algorithm is easily implemented. Examples from real data sets (credit-scoring and environmental statistics) illustrate the technique and the proposed methodology of minimum average variance estimation (MAVE).

Key words and phrases: Asymptotic distribution; Generalized partially linear model; Local linear smoother; Optimal consistency rate; Single-index model.

1. Introduction. Although the presence of nonlinearities in statistical data analysis is often modelled with non- and semi-parametric methods, there are still few noncritical semiparametric techniques. One argument that has been advanced is that - despite a reduction in dimensionality - the practical estimation still depends heavily on pilot estimators and complexity parameters. Another argument against finely tuned semiparametrics is that mathematical tools for inferential decisions and software implementations are either missing or not readily accessible. The purpose of this paper is to show that such critiques may be refuted even for the very flexible class of *Generalized Partially Linear Single Index Models* (GPLSIM):

$$y = \beta_0^T Z + g(\theta_0^T X) + \varepsilon, \tag{1.1}$$

where $E(\varepsilon|X,Z)=0$ almost surely, β_0 and θ_0 (with $\|\theta_0\|=1$) are unknown parameters, $g(\cdot)$ is an unknown link function. The GPLSIM (1.1) was first analyzed by Carroll *et al.* (1997) and contains the single-index models ($\beta_0 \equiv 0$), generalized partially linear models (X one dimensional and y observed logits), generalized linear models ($\beta_0 \equiv 0$ and g known) and of course the linear model (for $g \equiv 0$). Component identification of a more general model is investigated recently by Samarov *et al.* (2002). The advantage of the GPLSIM lies in its generality and its flexibility. The wide spread application of GPLSIMs though is somewhat obstructed by the facts described above: necessity of pilot estimators for θ_0 and complexity parameters such as bandwidths (to estimate the link function g).

The issue of the order of magnitude of the complexity parameter was addressed in Carroll *et al.* (1997, eqn.(18), p.483). The convenience of a root-n pilot estimator for θ_0 was employed in Härdle, Hall and Ichimura (1993) but was found to severely influence the final estimate. In practical application, these two important questions will be addressed in this paper: we will show that a simple multi-dimensional kernel estimator suffices to ensure root-n consistency of the parametric parts of (1.1) and that no under-smoothing is required for the proposed algorithm. In addition, we contribute to the theory of GPLSIMs by allowing the observations to be time series with weak mixing properties.

One motivation of our work comes from credit scoring and the study of nonlinear effects in retail banking. Another motivation comes from the analysis of circulatory and respiratory problems in Hong Kong and the study of the complicated effect of weather conditions on the health problems. *Credit Scoring* methods are designed to asses risk measures for potential borrowers, companies etc. Typically, the scoring is reduced to a classification or (quadratic) discriminant analysis problem, see Henley and Hand (1996) and Arminger *et al.* (1997). The credit data set of Müller and Rönz (2000) consists of 6180 cases with 8 metric variables $(x_2,\cdots,x_9)$ and 15 categorical explanatory variables $(x_{10},\cdots,x_{24})$. The response variable y was $=0$ or 1 on a rating scale $\{0,1\}$. There were 372 cases with a y value of 1. A scatterplot matrix of the observations (x_2,x_3,x_4,x_5) is given in Figure 1.

The distribution of the variable y (black points in Figure 1) shows a clear nonlinear structure and speaks therefore against a linear discriminant analysis. A logit model

$$\text{logit}\{P(y=1|X,Z)\} = \beta_0^T Z + \theta_0^T X \tag{1.2}$$

(also of linear structure) shows clear nonlinearity in the residuals, see Müller and Rönz (2000). Here X denotes the vector of metric variables and Z the vector of categorical variables. Müller and Rönz (2000) therefore applied a partially linear approach as in Severini and

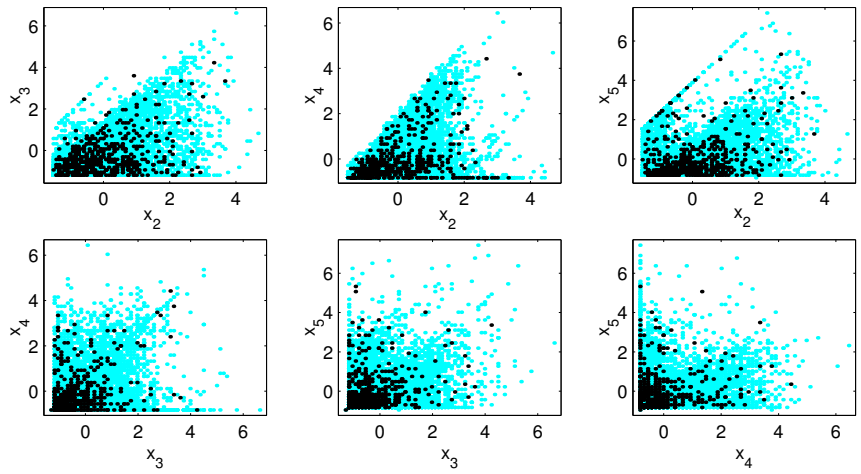

Figure 1: *Scatterplots: variables x_2 to x_5, observations corresponding to $y = 1$ are emphasized in black.*

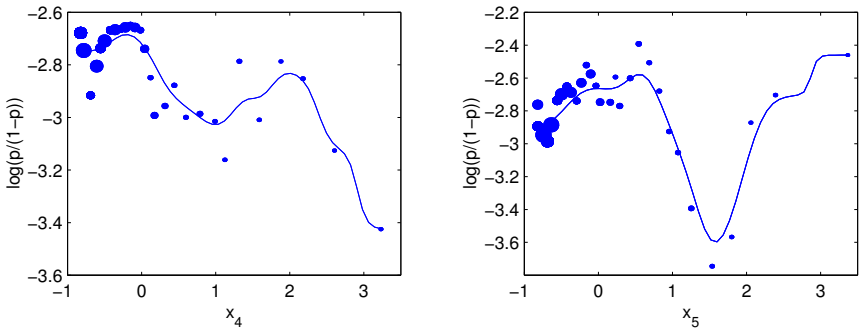

Figure 2: *Marginal dependency. Thicker bullets correspond to more observations in a class. The lines are local linear smoothers.*

Staniswalis (1994) by replacing one linear term in (1.2) operating on the metric variable x_5 by a nonparametric function $g(x_5)$ as shown in Figure 2.

We partition the range of x_4 (or x_5) into 50 intervals with equal lengths. We cluster the observations with x_4 (or x_5) in the same interval as one class. We calculate the relative frequencies $\hat{p}$ for $y = 1$. In Figure 2, the variable x_4 (or x_5) is plotted against the $\text{logit}(\hat{p}) = \log(\hat{p}/(1-\hat{p}))$. Using bootstrap, the nonlinearity was tested and found to be significant. The question of how to integrate further nonlinear influences by the other metric variables was analyzed in Müller and Rönz (2000) at a multidimensional kernel regression (e.g. on (x_4, x_5), see Figure 5.6 in their article) and found to be too difficult to implement due to the high dimensional kernel smoothing. The technique that we develop here will make it possible to overcome the dimensionality issue and indicate nonlinear influences on the logits via the GPLSIM.

The other motivation of this research comes from the investigation of the number of daily

hospital admissions of patients suffering from the circulatory and respiratory (CR) problems in Hong Kong from 1994-1996. There is a jump in the numbers at the beginning of 1995 due to the additional hospital beds released to accommodate CR patients from the beginning of 1995. We remove this jump by a simple kernel smoothing over time and denote the remaining time series by y_t. The pollutants and weather conditions might cause the CR problems. The pollutants include sulphur dioxide (x_{1t}, in μgm^{-3}), nitrogen dioxide (x_{2t}, in μgm^{-3}), respirable suspended particulates (x_{3t}, in μgm^{-3}) and ozone (x_{4t}, in μgm^{-3}), and weather conditions include temperature (x_{5t}, in ^oC) and relative humidity (x_{6t}, in %). It is obvious that the higher the levels of air pollutants are, the stronger they tend to cause health problems. Furthermore, simple kernel smoothing suggests that we can approximate the relations between y_t and the pollution levels linearly; see Figure 3. However, for the other covariates such as temperature and humidity, the relations are unknown and might be nonlinear. Figure 3 is simple regression analyses based on kernel smoothing. The relation of y_t with NO_2 is almost linear, but the relation of y_t with humidity is nonlinear and hard to explain. To explore the relation between y_t and air pollutants and weather conditions, we may consider the following model

$$y_t = \beta^T Z_t + g(\theta^T X_t) + \varepsilon_t, \tag{1.3}$$

where Z_t consists of levels of pollutants and their lagged variables, and X_t consists of weather conditions and their lagged variables.

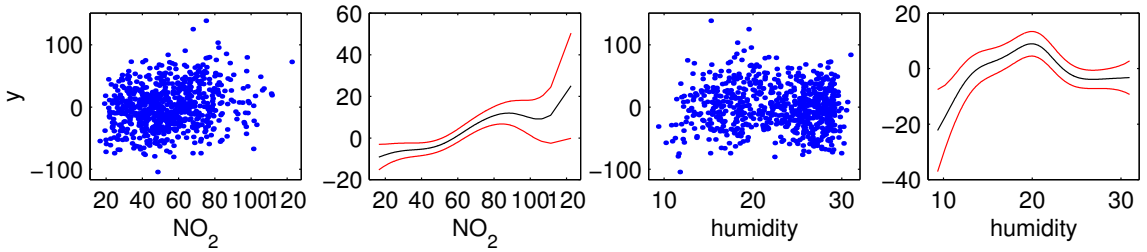

Figure 3: *The first and third panels are the plots of daily y against NO_2 and humidity respectively. In the second and fourth panels, the central lines are kernel smoothers of y on NO_2 and humidity respectively, the upper and lower lines are the corresponding 95% pointwise confidence intervals.*

Before we present our estimation method, we briefly summarize the current four main critiques on the estimations of model (1.1) or its special cases. (1) *Heavy computational burden:* see, for example, Härdle *et al.*, Carroll *et al.* (1997), Xia and Li (1999) and Xia *et al.* (1999). These methods include complicated optimization techniques and no simple algorithm is available up to now. (2) *Strong restrictions on link functions or design of covariates:* Li (1991) required symmetric distribution of the covariate; Härdle and Stoker

(1989) and Hristache *et al.* (2001a) required that $|Eg'(\theta_0^T X)|$ is away from 0. If these conditions are violated, their methods cannot obtain useful estimators. (3) *Inefficiency:* The method of Härdle and Stoker (1989) and the method of Hristache *et al.* (2001a, 2001b) are not asymptotically efficient in the semi-parametric sense. (4) *Under-smoothing:* Most of the methods mentioned above require a bandwidth that is much smaller than the data-driven bandwidth in order to allow the estimator of the parameters to achieve root-n consistency, i.e. under-smoothing the link function is needed; see, Härdle and Stoker (1989) and Hristache *et al.* (2001a, 2001b), Hall (1989) and Carroll *et al.* (1997) among others. More discussions on the selection of bandwidth for the partially linear model can be found in Linton (1995). In this paper we present the minimum average variance estimation (MAVE) method that will provide a remedy to these four weak points.

2. Estimation method. The basic algorithm for estimating the parameters in (1.1) is based on observing that

$$(\beta_0, \theta_0) = \arg\min_{\beta,\theta} E\Big[y - \{\beta^T Z + g(\theta^T X)\}\Big]^2 \tag{2.1}$$

subject to $\theta^T\theta = 1$. By conditioning on $\xi = \theta^T X$, we see that (2.1) equals $E_\xi \sigma^2_{\beta,\theta}(\xi)$ where

$$\sigma^2_{\beta,\theta}(\xi) = E\Big[\Big(y - \{\beta^T Z + g(\xi)\}\Big)^2 \Big| \theta^T X = \xi\Big].$$

It follows that

$$E\Big[y - \{\beta^T Z + g(\theta^T X)\}\Big]^2 = E_\xi \sigma^2_{\beta,\theta}(\theta^T X).$$

Therefore, minimization (2.1) is equivalent to ,

$$(\beta_0, \theta_0) = \arg\min_{\beta,\theta} E_\xi \sigma^2_{\beta,\theta}(\xi) \tag{2.2}$$

subject to $\theta^T\theta = 1$. Let $\{(X_i, Z_i, y_i)\ i = 1, 2, \cdots, n\}$ be a sample from (X, Z, y). The conditional expectation in (2.2) is now approximated by the sample analogue. For X_i close to x, we have the following local linear approximation

$$y_i - \beta_0^T Z - g(\theta_0^T X_i) \approx y_i - \beta_0^T Z_i - g(\theta_0^T x) - g'(\theta_0^T x) X_{i0}^T \theta_0,$$

where $X_{i0} = X_i - x$. Following the idea of local linear smoothing, we may estimate $\sigma^2_{\beta,\theta}(\theta^T x)$ by

$$\hat\sigma^2_{\beta,\theta}(\theta^T x) = \min_{a,d} \sum_{i=1}^n \Big\{y_i - \beta^T Z_i - a - d X_{i0}^T \theta\Big\}^2 w_{i0}. \tag{2.3}$$

Here, $w_{i0} \geq 0, i = 1, 2, \cdots, n$, are some weights with $\sum_{i=1}^n w_{i0} = 1$, typically centering at x. Let $X_{ij} = X_i - X_j$. By (2.2) and (2.3), our estimation procedure is to minimize

$$\frac{1}{n}\sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} \left\{ y_i - \beta^T Z_i - a_j - d_j X_{ij}^T \theta \right\}^2 w_{ij} \tag{2.4}$$

with respect to (a_j, d_j) and (β, θ), where $G(\cdot)$ is another weight function that controls the contribution of (X_j, Z_j, y_j) to the estimation of β and θ. For example, when the model is assumed to be heteroscedastic and $Var(y|X, Z) = V(\theta_0^T X)$, then $G(.) = V(.)$; see Härdle *et al.* (1993) and Carroll *et al.* (1997). $I_n(x)$ is employed here for technical purpose to handle the boundary points. It is given in the next section. See also Härdle *et al.* (1993). For simplicity, we can take $I_n(\cdot) = 1$ in practice. We call the estimation procedure the minimum average (conditional) variance estimation (MAVE) method. Minimizing (2.4) is a typical quadratic programming and can be solved easily. Next, we give a GPLSIM algorithm. Given (β, θ), we have

$$\begin{pmatrix} a_j \\ d_j \end{pmatrix} = \left\{ \sum_{i=1}^{n} w_{ij} \begin{pmatrix} 1 \\ X_{ij}^T\theta \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij}^T\theta \end{pmatrix}^T \right\}^{-1} \sum_{i=1}^{n} w_{ij} \begin{pmatrix} 1 \\ X_{ij}^T\theta \end{pmatrix} (y_i - \beta^T Z_i). \tag{2.5}$$

Given (a_j, d_j), we calculate

$$\begin{pmatrix} \beta \\ \theta \end{pmatrix} = \left\{ \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} w_{ij} \begin{pmatrix} Z_i \\ d_j X_{ij} \end{pmatrix} \begin{pmatrix} Z_i \\ d_j X_{ij} \end{pmatrix}^T \right\}^{-1} \times \sum_{j=1}^{n} G(\theta^T X_j) I_n(X_j) \sum_{i=1}^{n} w_{ij} \begin{pmatrix} Z_i \\ d_j X_{ij} \end{pmatrix} (y_i - a_j) \tag{2.6}$$

and standardize $\theta := \theta/|\theta|$. Here and later, $|\gamma| = (\gamma^T\gamma)^{1/2}$ for any vector γ. The minimization in (2.4) can be solved by iterations between (2.5) and (2.6).

The choice of the weights w_{ij} plays an important role in different estimation methods. See Xia *et al.* (2002) and Hristache *et al.* (2001a, 2001b). In this paper, we use two sets of weights. Suppose $H(\cdot)$ and $K(\cdot)$ are a p-variate and a univariate density function respectively. The first set of weights is $w_{ij} = H_{b,i}(X_j)/\sum_{\ell=1}^n H_{b,\ell}(X_j)$, where $H_{b,i}(X_j) = b^{-p}H(X_{ij}/b)$ and b is a bandwidth. This is a multivariate dimensional kernel weight. For this kind of weights, we set $I_n(x) = 1$ if $n^{-1}\sum_{\ell=1}^n H_{b,\ell}(x) > c_0$; 0 otherwise for some constant $c_0 > 0$. Iterating (2.5) and (2.6) until convergence, denote the estimators (i.e., the final values) of θ and β by $\tilde{\theta}$ and $\tilde{\beta}$ respectively. Because of the so-called "curse of dimensionality", the estimation based on this kind of weights is not efficient. However, the multivariate kernel

weight can help us to find an appropriate initial step of the estimation. We then use single-index kernel weights $w_{i,j}^{\theta} = K_{h,i}^{\theta}(\theta^T X_j)/\sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(\theta^T X_j)$, where $K_{h,i}^{\theta}(v) = h^{-1}K\{(\theta^T X_i - v)/h\}$, h is the bandwidth and θ is the previous estimate of θ_0. Here, we take $I_n(x) = 1$ if $n^{-1}\sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(\theta^T x) > c_0$; 0 otherwise. Iterating (2.5) and (2.6) until convergence, denote the estimators (i.e. the final values) of θ and β by $\hat{\theta}$ and $\hat{\beta}$ respectively. After obtaining estimates $\hat{\theta}$ and $\hat{\beta}$, we can then estimate $g(v)$ by the solution of a_j in (2.5) with $\theta^T X_j$ replaced by v, denote the estimate by $\hat{g}(v)$. A computer code for the above algorithm is available at $http://www.hku.hk/statisics/paper$.

The main results of this algorithm are: (1) A $\sqrt{n}$-consistent pilot estimator is not needed, see Theorem 1 below. This solves the problems addressed in Carroll *et al.* (1997); (2) Convergence of the GPLSIM algorithm is proved, see the proof of Theorem 1 in section 6; (3) An "undersmooth bandwidth" is not needed, since the cross-validated bandwidth (for a smoothing estimate of g) suffices. This makes the algorithm stable and frees it from the audible critique on "the necessity of uncontrollable hyperparameters"; (4) Under some assumptions, the estimators of the parameters is asymptotically efficient in semi-parametric sense, see Carroll *et al.* (1997); and (5) The GPLSIM algorithm is applicable to time series. This feature makes the technique widely applicable in nonlinear time series analysis.

Let $U = (X^T, Z^T)^T$. Suppose $\{(U_i, y_i), i = 1, \cdots, n\}$ is a set of observations. We make the following assumptions on the stochastic nature of the observations, the link function and the kernel functions.

(C1) The observations are a strongly mixing and stationary sequence with geometric decaying mixing rate $\alpha(k)$.

(C2) With Probability 1, X lies in a compact set $\mathcal{D}$; the marginal density functions f of X and f_θ of $\theta^T X$ for any $|\theta| = 1$ have bounded derivatives; regions $\{x : f(x) \geq c_0\}$ and $\{x : f_\theta(\theta^T x) > c_0\}$ for all $\theta : |\theta| = 1$ are non-empty.

(C3) For any perpendicular unit norm vectors θ and ϑ, the joint density function $f(u_1, u_2)$ of $(\theta^T X, \vartheta^T X)$ satisfies $f(u_1, u_2) < c f_{\theta^T X}(u_1) f_{\vartheta^T X}(u_2)$, where c is a constant.

(C4) g has bounded, continuous third order derivative; the conditional expectations $E(Z|X = x)$, $E(ZZ^T|X = x)$, $E(U|\theta^T X = v)$ and $E(UU^T|\theta^T X = v)$ have bounded derivatives; $E(y^r|X = x)$, $E(|Z|^r|X = x)$, $E(|Z_\ell||Z_1|\Big|X_1 = x_1, X_\ell = x_\ell)$ and $E(|Z_\ell||Z_1|\Big|\theta^T X_1 = u, \theta^T X_\ell = v)$ are bounded by a constant for all $\ell > 0$, x_1, x_ℓ, x, u and v, where $r > 2$.

(C5) H is a density function with bounded derivative and compact support $\{|x| \leq a_0\}$ for some $a_0 > 0$; K is a symmetric density function with bounded derivative and compact support $[-b_0, b_0]$ for some $b_0 > 0$ and that the Fourier transform of K is absolute integrable.

(C6) $E\{(Z - E(Z|X))(Z - E(Z|X))^T\}$ is a positive definite matrix.

The mixing rate in (C1) can be relaxed to algebraic rate $\alpha(k) = O(k^{-\rho})$. Suppose the bandwidth $h \sim n^{-\delta}$. Then the mixing rate satisfying the following equation is sufficient.

$$\sum_{n=1}^{\infty} n^{-\{\frac{1}{2}-\frac{1}{r}-\delta(\frac{1}{2}+\frac{1}{r})\}\rho+2p+1+\frac{1}{r}+(\frac{1}{2}+\frac{1}{r})\delta}(\log n)^{\rho/2} < \infty.$$

The regions with positive densities in (C2) are needed to avoid zero values of the denominator of kernel estimator of regression. There are different approaches for this purpose. See, e.g. Härdle *et al.* (1993), Härdle and Stoker (1989) and Linton (1995). However, their ideas are the same. We can further assume that c_0 decreases to 0 with n at a slow speed, but it makes no difference in practices. Assumption (C3) ensures successful searching for the direction θ globally. If we restrict the searching area, the assumption can be removed; see Härdle *et al.* (1993). The third order derivatives in (C4) is needed for higher order expansion. Actually, existence of second order derivative is sufficient for the root-n consistency. In this paper, we only employ kernel functions with compact support as in (C5). (C6) is imposed for identification. Similarly, if we search for the direction θ in a small neighbour of θ_0 as in Härdle *et al.* (1993) and Carroll *et al.* (1997), (C6) can be removed.

Lemma 1. *Let $\tilde{\beta}$ and $\tilde{\theta}$ be the estimators based on the multi-dimensional kernel weight. Suppose that (C1)-(C6) hold, $b \to 0$ and $nb^{p+2}/\log n \to \infty$. If we start the estimation procedure with θ such that $\theta^T\theta_0 \neq 0$, then*

$$\tilde{\theta} - (\pm\theta_0) = o_P(1), \qquad \tilde{\beta} - \beta_0 = o_P(1),$$

where the sign before θ_0 is determined in accordance with the sign of $\theta^T\theta_0$.

Let $\mu_\theta(x) = E(X|\theta^T X = \theta^T x)$, $\nu_\theta(x) = E(Z|\theta^T X = \theta^T x)$, and for $k = 0$ and 2,

$$W_k = E\Big\{G(\theta_0^T X)I(f_{\theta_0}(\theta_0^T X) > c_0) \begin{pmatrix} Z - \nu_{\theta_0}(X) \\ \{\pm g'(\theta_0^T X)\}\{X - \mu_{\theta_0}(X)\} \end{pmatrix} \times \begin{pmatrix} Z - \nu_{\theta_0}(X) \\ \{\pm g'(\theta_0^T X)\}\{X - \mu_{\theta_0}(X)\} \end{pmatrix}^T |\varepsilon|^k \Big\}.$$

Theorem 1. *Let $(\hat{\beta}, \hat{\theta})$ be the estimators based on the single-index kernel weight starting with $(\beta, \theta) = (\tilde{\beta}, \tilde{\theta})$. Suppose (C1)-(C6) hold, $h \sim n^{-\delta}$ with $1/6 < \delta < \min(1/4, 1 - 2/r)$ and that $E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} = 0$ almost surely. Then*

$$n^{1/2} \begin{pmatrix} \hat{\beta} - \beta_0 \\ \hat{\theta} - (\pm\theta_0) \end{pmatrix} \xrightarrow{D} N(0, W_0^- W_2 W_0^-),$$

where W_0^- is the Moore-Penrose inverse of W_0, the signs before θ_0 and g' (in $W_k, k = 0, 2$) are determined in accordance with the sign of $\tilde{\theta}^T\theta_0$. If further the density function $f_{\theta_0}(v)$ of $\theta_0^T X$ is positive and the derivative of $E(\varepsilon^2|\theta_0^T X = v)$ exists, then

$$(nh)^{1/2}\{\hat{g}(v) - g(v) - \frac{1}{2}\kappa_2 g''(v)h^2\} \xrightarrow{D} N(0, f_{\theta_0}^{-1}(v)\int (K(v))^2 dv E(\varepsilon^2|\theta_0^T X = v)),$$

where $\kappa_2 = \int K(v)v^2 dv$.

If $E\{\varepsilon_i|(X_j, Z_j, y_j), j < i\} \neq 0$, then the asymptotic normal distribution still holds, but the variance matrix in the distribution depends on the structure of the stochastic process of the observations. If $E(\varepsilon^2|X, Z) = \sigma^2$ is constant, then the asymptotic distribution of $(\hat{\beta}, \hat{\theta})$ is the same as that obtained by Carroll *et al.* (1997). They further showed that their estimator is efficient in the semiparametric sense under some mild conditions. Therefore our estimator is also efficient in the semiparametric sense under the same conditions. Bandwidth selection is always an important issue for nonparametric methods. One of the advantages of our method is that we don't need under-smoothing the link function when $r > 2.5$. Therefore, most commonly used bandwidth selection methods can be employed here. Consider estimation of g, i.e. a_j (and d_j), at the final step of the iterations. For a given function $w(\cdot)$ with compact support, minimizing the asymptotic weighted mean squared error with weight $f_{\theta_0}(\cdot)w(\cdot)$ yields the optimal global bandwidth

$$h_o = \left\{\frac{\sigma^2 \int w(u)du \int (K(u))^2 du}{\kappa_2^2 \int g''(u) f_{\theta_0}(u) w(u) du}\right\}^{1/5} n^{-1/5}.$$

See also the discussion in Carroll *et al.* (1997). Both the cross-validation bandwidth selection method and the plug-in method can be used to obtain bandwidths that are asymptotically consistent of h_o.

4. Numerical Comparisons. In this section, we first use an example to demonstrate the relation between estimation errors and the bandwidth. We then use the examples in Härdle *et al.* (1993) and Carroll *et al.* (1997) to check the performance of our estimation method for finite data sets. In our simulations, kernel functions $H(x) = 3(1-|x|^2)I(|x| < 1)/4$ and $K(u) = 3(1-u^2)I(|u| < 1)/4$ are used.

Example 4.1. Consider the following model

$$y_t = \beta_{01} z_{1t} + \beta_{02} z_{2t} + 2\exp\{-3(\theta_{01}x_{t-1} + \theta_{02}x_{t-2} + \theta_{03}x_{t-3})^2\} + 0.5\varepsilon_t,$$

where $x_t = 0.4x_{t-1} - 0.5x_{t-2} + u_t$ with $u_t, t = 1, 2, \cdots, \overset{IID}{\sim} Uniform(-1, 1)$; z_{1t} and $z_{2t}, t = 1, 2, \cdots$ are IID as binary distribution taking values 0 and 1 with probability 0.5 each; $\varepsilon_t, t =$

$1, 2, \cdots, \overset{IID}{\sim} N(0,1)$; and that $\{u_t\}, \{z_{1t}\}, \{z_{1t}\}$ and $\{\varepsilon_t\}$ are independent series. Here, $Z_t = (z_{1t}, z_{2t})^T$ and $X_t = (x_{t-1}, x_{t-2}, x_{t-3}, \cdots, x_{t-p})^T$. The true parameters are $\beta = (\beta_{01}, \beta_{02})^T = (1,2)^T$ and $\theta = (\theta_{01}, \theta_{02}, \theta_{03}, \cdots, \theta_{0p})^T = (-2/3, 1/3, 2/3, 0, \cdots, 0)^T$. We define the estimation errors as $e_\beta = (|\hat{\beta}_1 - \beta_{01}| + |\hat{\beta}_2 - \beta_{02}|)/2$ and $e_\theta = 1 - |\hat{\theta}^T \theta_0|$ for $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$ and $\hat{\theta}$ respectively. With different dimension p, sample sizes and bandwidths, the logarithm of the average errors (the solid lines) are shown in Figure 4 (the number of replications is 100). The vertical lines are the corresponding average of cross-validation bandwidths. Figure 4 shows that the estimation procedure works quite well and the cross-validation bandwidth is applicable to the estimation of the parameters.

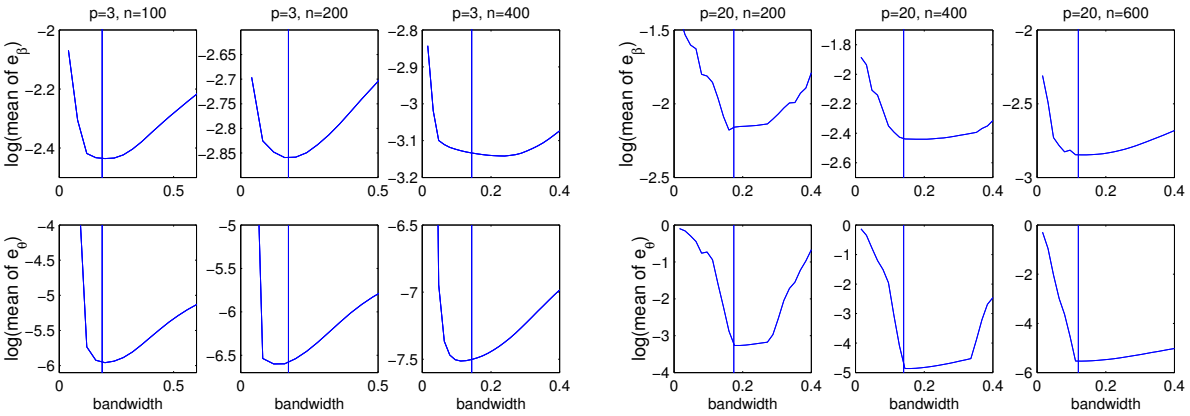

Figure 4: *Simulation results of Example 4.1. The solid lines are logarithms of the means of the estimation errors from 100 replications; The vertical lines are means of corresponding cross-validation bandwidths.*

Example 4.2. Consider the following two models

$$y = 4\{(x_1 + x_2 - 1)/\sqrt{2}\}^2 + 4 + 0.2\varepsilon, \tag{4.1}$$

$$y = \sin\{\pi((x_1 + x_2 + x_3)/\sqrt{3} - A)/(B - A)\} + \beta Z + 0.1\varepsilon, \tag{4.2}$$

where x_1, x_2, x_3 are independent uniformly distributed on $[0,1]$, $A = 0.3912$ and $B = 1.3409$. Model (4.1) was used by Härdle *et al.* (1993), in which $\theta_0 = (\theta_{11}, \theta_{12})^T = (1/\sqrt{2}, 1/\sqrt{2})^T$. Model (4.2) was used by Carroll *et al.* (1997), in which $\theta_0 = (\theta_{21}, \theta_{22}, \theta_{23})^T = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T$. We start the simulation for model (4.1) with $\theta = (1,3)^T/\sqrt{10}$ and model (4.2) with $\theta = (0,1,2)^T/\sqrt{5}$. The cross-validation bandwidth is used. The number of replications is 100. With sample size $n = 50, 100$ and 200, the simulation results are listed in Table 1.

For model (4.1), the corresponding simulation results of $\phi = \arccos(\theta_{11})$ were 0.766(0.103), 0.792(0.084), 0.782(0.045) for n =50, 100 and 200 respectively in of Härdle *et al.* (1993).

Our results outperform theirs. A possible reason is that minimizing the cross-validation type of residuals was used to estimate the parameters in their paper, which reduces the estimation efficiency. See Xia *et al.* (1999) for details. For model (4.2), the corresponding simulation results of Carroll *et al.* (1997) for θ_{21}, θ_{22} and θ_{23} were (1.4e-4), (1.6e-4) and (1.3e-4) respectively when $n = 200$. Our results also improve theirs.

TABLE 1: Mean and mean squared error (in parentheses) of the estimated parameters for models (4.1) and (4.2)

n	Model (4.1)			Model (4.2)			
	θ_{11}	θ_{12}	$\phi = \arccos(\theta_{11})$	θ_{21}	θ_{22}	θ_{23}	β
50	0.7117	0.6965	0.7746	0.5793	0.5727	0.5785	0.2967
	(0.0040)	(0.0045)	(0.0918)	(5.5e-4)	(5.7e-4)	(6.5e-4)	(1.1e-3)
100	0.7074	0.7047	0.7835	0.5785	0.5780	0.5748	0.2972
	(0.0015)	(0.0015)	(0.0541)	(2.8e-4)	(2.6e-4)	(2.2e-4)	(4.7e-4)
200	0.7071	0.7059	0.7845	0.5776	0.5770	0.5772	0.2992
	(0.0008)	(0.0008)	(0.0403)	(1.2e-4)	(1.3e-4)	(1.2e-4)	(2.5e-4)

5. Real Data Analysis. Now we return to our real data sets in section 1. The Epanechnikov kernel and the cross-validation bandwidths are used in the calculations.

Credit Scoring. We consider model (1.1) with all the covariates by taking $Z = (x_2, x_3, x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21}, x_{22}, x_{23}, x_{24})^T$, $X = (x_4, x_5)^T$ and assume $E(\varepsilon^2|X, Z) = \sigma^2$ is a constant. Here, x_4 and x_5 are standardized respectively for ease of calculations. Applying the estimation procedure to the data set, we obtain the estimates of the parameters as listed in table 2. See Müller and Rözn (2000) for more explanations. The estimate of the unknown function is shown in the right panel of Figure 5. The nonlinearity in x_4 and x_5, i.e. $\hat{\theta}^T X = 0.249x_4 + 0.969x_5$, is clear as shown in Figure 5.

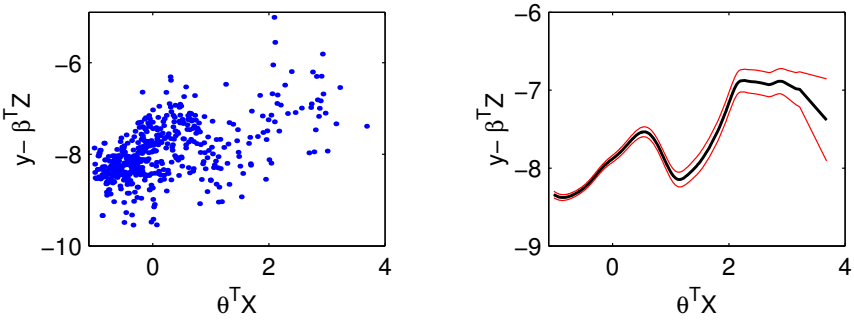

Figure 5: *Estimation results of the credit scoring data. The left panel is $y - \hat{\beta}^T Z$ plotted against $\hat{\theta}^T X$. The right panel is the estimated g and 95% symmetric pointwise confidence interval.*

Table 2. Estimation results of the Credit scoring data

variable	coeff.	S.E.	variable	coeff.	S.E.	variable	coeff.	S.E.
2	0.159	0.163	17#2	**-1.718**	0.472	20#3	-0.082	0.294
3	0.021	0.114	17#3	**-1.211**	0.433	20#4	0.263	0.251
6	-0.109	0.105	17#4	**1.977**	0.576	21#2	**-2.194**	0.683
7	**-0.454**	0.119	17#5	**4.715**	0.932	21#3	**-1.490**	0.363
8	0.189	0.120	17#6	**-1.327**	0.316	22#2	-1.102	0.582
9	0.032	0.091	18#2	**2.145**	0.528	22#3	-0.785	0.490
10#2	**0.817**	0.302	18#3	**1.037**	0.413	22#4	0.753	0.715
11#2	0.188	0.293	18#4	0.878	0.447	22#5	0.770	0.584
12#2	**0.635**	0.303	18#5	**1.756**	0.359	22#6	**-3.837**	0.957
13#2	**-0.815**	0.276	18#6	**1.876**	0.449	22#7	**2.253**	0.608
14#2	**1.680**	0.544	18#7	**1.770**	0.551	22#8	0.838	0.531
15#2	**1.416**	0.347	19#2	0.416	0.369	22#9	**1.441**	0.526
15#3	**2.411**	0.469	19#3	**1.287**	0.307	22#10	-1.519	1.199
15#4	**3.247**	0.520	19#4	-0.966	0.539	22#11	-0.644	0.510
15#5	**2.782**	0.617	19#5	1.343	0.673	23#2	0.087	0.350
15#6	**0.987**	0.374	19#6	**1.691**	0.465	23#3	0.787	0.499
16#2	0.214	0.476	19#7	0.992	0.539	24#2	**1.717**	0.612
16#3	0.680	0.431	19#8	**-1.170**	0.566	4	**0.249**	0.026
16#4	**1.714**	0.539	19#9	0.173	0.608	5	**0.969**	0.007
16#5	**1.218**	0.442	19#10	**1.070**	0.348			
16#6	**1.588**	0.465	20#2	**2.021**	0.539	$\hat{\sigma}^2 = 0.1589$		

Circulatory and respiratory problems in Hong Kong. Due to the hospital booking system, the day-of-the-week can affect y_t. We use dummy variables to describe the day of the t'th observation by a 6-dimension vector $(D_{t1}, \cdots, D_{t6})$, where $D_{tk} = 1$ if the observation is taken on the k'th day of a week; 0 otherwise. Together with lagged variables of pollutants and weather conditions in one week, we take $Z_t = (D_{t1}, \cdots, D_{t6}, x_{1,t-1}, \cdots, x_{1,t-7}, x_{2,t-1}, \cdots, x_{2,t-7}, x_{3,t-1}, \cdots, x_{3,t-7}, x_{4,t-1}, \cdots, x_{4,t-7})^T$ and $X_t = (x_{5,t-1}, \cdots, x_{5,t-7}, x_{6,t-1}, \cdots, x_{6,t-7})^T$ in model (1.1). Here, $x_{1,t}, \cdots, x_{6,t}$ are standardized. We further assume $E(\varepsilon_t^2 | X_t, Z_t) = \sigma^2$ is a constant. By the asymptotic distribution of the parameters, we remove the covariate with smallest t-values in the estimated model one by one and re-estimate the model. Continue this procedure until all the covariates have t-values larger than 1.8. We finally obtain the following model (the values in the parentheses are the corresponding standard errors of the estimators)

$$\begin{aligned} y_t = &-0.3831 D_{t1} - 0.1728 D_{t2} - 0.5636 D_{t3} - 0.7399 D_{t4} - 1.0871 D_{t5} - 1.1562 D_{t6} \\ &\quad (0.0942) \quad (0.0943) \quad (0.0945) \quad (0.0947) \quad (0.0946) \quad (0.0942) \\ &+ 0.0957 x_{2,t-1} + \hat{g}(0.4257 x_{5,t-2} - 0.6079 x_{5,t-5} + 0.4974 x_{6,t-4} + 0.4492 x_{6,t-7}). \\ &\quad (0.0287) \qquad (0.1576) \qquad (0.1104) \qquad (0.1745) \qquad (0.1475) \end{aligned}$$

The estimated link function $\hat{g}$ is shown in Figure 6.

Based on this model, the effects of weather conditions on the CR problems are as follows. The coefficients of temperatures $x_{5,t-2}$ and $x_{5,t-5}$ forms a contrast. Together with Figure 6, it suggests that a rapid temperature variation (rather than the temperature itself) will increase the hospital admission y_t. The coefficients of humidity $x_{6,t-4}$ and $x_{6,t-7}$ have about the same value, which can be taken as an average. Together with Figure 6, it suggests that extreme dry or wet weather will increase the hospital admission in Hong Kong.

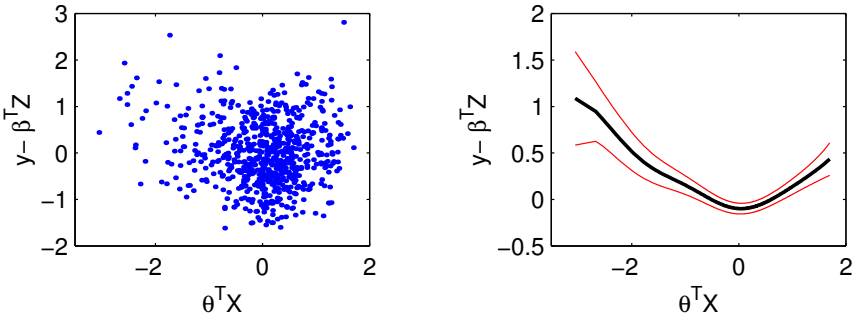

Figure 6: *Estimation results for the circulatory and respiratory problems in Hong Kong. The left panel is $y_t - \hat{\beta}^T Z_t$ plotted against $\hat{\theta}^T X_t$. The right panel is the estimated g and 95% symmetric pointwise confidence interval.*

6. Proofs. The basic tools are given in Lemmas A.1-A.3. Some simple calculation results are listed in Lemmas A.4-A.6. Based on these Lemmas, Lemma 1 and Theorem 1 are proved. For simplicity, we shall prove Theorem 1 for the case $\tilde{\theta}^T\theta_0 > 0$. The proof of Theorem 1 for the other case ($\tilde{\theta}^T\theta_0 < 0$) is similar. Some differences in the proofs between these two cases are addressed in the context. Let $\delta_\theta = |\theta - \theta_0|$, $\delta_\beta = |\beta - \beta_0|$ and $\delta_\gamma = \delta_\theta + \delta_\beta$. In a bounded parameter space, δ_θ, δ_β and δ_γ are bounded. Let $\delta_{pn} = \{\log n/(nb^p)\}^{1/2}$, $\tau_{pn} = b^2 + \delta_{pn}$, $\delta_n = \{\log n/(nh)\}^{1/2}$, $\tau_n = h^2 + \delta_n$ and $\delta_{0n} = (\log n/n)^{1/2}$. By the condition $h \sim n^{-\delta}$ with $1/6 < \delta < 1/4$, we have $\delta_{0n} \ll h^2 \ll h^{-1}\delta_n$ and $\delta_n \ll h$. We shall use these relations frequently in our calculations. Let $\Theta = \{\theta : |\theta| = 1\}$. Suppose A_n is a matrix. $A_n = O(a_n)$ (or $o(a_n)$) means every element in A_n is $O(a_n)$ (or $o(a_n)$) almost surely. We adopt the consistency in the sense of "almost surely" because we need to prove the convergence of the algorithm, which theoretically need infinite iterations. Let $c, c_1, c_2, \cdots$ be a set of constants. For ease of exposition, c may have different values at different places. We abbreviate $K_h(\theta^T X_{i0})$ and $H_b(X_{i0})$ as $K^\theta_{h,i}(x)$ (or $K^\theta_{h,i}$) and $H_{b,i}(x)$ (or $H_{b,i}$) respectively in the following context. We take $G(\cdot) \equiv 1$ in the proofs for simplicity. We further assume that $\kappa_2 \stackrel{def}{=} \int K(v)v^2 = 1$ and $\mathcal{H}_2 \stackrel{def}{=} \int H(U)UU^T dU = I_{p\times p}$; otherwise we may take $K(v) =: K(v/\sqrt{\kappa_2})/\sqrt{\kappa_2}$ and $H(U) =: H(\mathcal{H}_2^{-1/2}U)(det(\mathcal{H}_2))^{-1/2}$.

Lemma A.1. *Suppose that* $m_1(\theta, x, z)$ *and* $\varphi(x, z, v)$ *are measurable functions with* $\sup_{\theta\in\Theta} E|m_1(\theta,\ X,\ Z)|^r < \infty$ *for some* $r > 2$ *and* $\sup_{x,z} |m_1(\theta, x, z)-m_1(\theta_0, x, z)| < c|\theta-\theta_0|$. *Let* $\varphi_i = \varphi(X_i, Z_i, y_i)$. *Assume* $\sup_{\theta\in\Theta,v} E(|\varphi_i|^r \Big|\theta^T X = v) < \infty$ *and* $\sup_{\theta\in\Theta,u,v} E(|\varphi_i \varphi_1|\Big|\theta^T X_1 = u, \theta^T X_i = v) < c$ *for all* $i > 1$. *Let* $g(v)$ *be any function with continuous second order derivative,* $m(u, v) = g(u) - g(v) - g'(v)(u - v) - g''(v)(u - v)^2/2$ *and* $\zeta_i^{k,\ell} = m(\theta_0^T X_i, \theta_0^T x)\mathbf{z}_i^k(\theta^T X_{i0})^\ell$ *where* $\mathbf{z}_i$ *is any component of* Z_i, $k = 0, 1$ *and* $\ell = 0, 1$. *If (C1) hold, then*

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^n\{m_1(\theta, X_i, Z_i) - Em_1(\theta, X_i, Z_i)\}\Big| = O(\delta_{0n}),$$

$$\sup_{|\theta-\theta_0|<a_n}\Big|\frac{1}{n}\sum_{i=1}^n\{m_1(\theta, X_i, Z_i) - m_1(\theta_0, X_i, Z_i)\}\varepsilon_i\Big| = O(a_n\delta_{0n}),$$

where $a_n \to 0$ *as* $n \to \infty$. *If further (C2)-(C5) hold,* $h \sim n^{-\delta}$ *with* $0 < \delta < 1 - 2/r$, *then*

$$\sup_{x\in\mathcal{D}}\Big|\frac{1}{n}\sum_{i=1}^n\{H_{b,i}\varphi_i - E(H_{b,i}\varphi_i)\}\Big| = O(\delta_{pn}), \quad \sup_{\substack{\theta\in\Theta\\ x\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^n\{K_{h,i}^\theta\varphi_i - E(K_{h,i}^\theta\varphi_i)\}\Big| = O(\delta_n),$$

$$\sup_{\substack{|\theta-\theta_0|<a_n\\ x\in\mathcal{D}}}\Big|\frac{1}{n}\sum_{i=1}^n\{K_{h,i}^\theta\zeta_i^{k,\ell} - E(K_{h,i}^\theta\zeta_i^{k,\ell})\}\Big| = O\{\delta_n h^\ell(a_n^2 + h^2)\}.$$

Proof. The proofs of Lemma A.1 are quite standard; see, e.g. Härdle *et al.* (1988) and Xia and Li (1999). We here give the details for the last two equations. Note that $\Theta\otimes\mathcal{D} \subset \mathbb{R}^{2p}$ is bounded. There are n^{2p} balls B_{n_k} centered at (θ_{n_k}, x_{n_k}), $1 \le k \le n^{2p}$, with diameter less then $cn^{-1/2}h^{3/2}(> c/n)$, such that $\Theta \otimes \mathcal{D} \subset \cup_{1\le k\le n^{2p}} B_{n_k}$. Then

$$\begin{aligned}
&\sup_{x\in\mathcal{D},\theta\in\Theta}\Big|\frac{1}{n}\sum_{i=1}^n\{K_{h,i}^\theta(x)\varphi_i - E(K_{h,i}^\theta(x)\varphi_i)\}\Big| \\
&\le \max_{1\le k\le n^{2p}}\Big|\frac{1}{n}\sum_{i=1}^n\Big[K_{h,i}^{\theta_{n_k}}(x_{n_k})\varphi_i - E\{K_{h,i}^{\theta_{n_k}}(x_{n_k})\varphi_i\}\Big]\Big| \\
&\quad + \max_{1\le k\le n^{2p}}\sup_{(\theta,x)\in B_{n_k}}\Big|\frac{1}{n}\sum_{i=1}^n\Big[\{K_{h,i}^\theta(x) - K_{h,i}^{\theta_{n_k}}(x_{n_k})\}\varphi_i \\
&\qquad\qquad -E\{(K_{h,i}^\theta(x) - K_{h,i}^{\theta_{n_k}}(x_{n_k}))\varphi_i\}\Big]\Big| \\
&\stackrel{def}{=} \max_{1\le k\le n^{2p}}|R_{n,k,1}| + \max_{1\le k\le n^{2p}}\sup_{(\theta,x)\in B_{n_k}}|R_{n,k,2}|. \qquad (6.1)
\end{aligned}$$

By assumption (C5), we have

$$\begin{aligned}
\max_{\substack{1\le k\le n^{2p}\\ x\in\mathcal{D}}}\sup_{(\theta,x)\in B_{n_k}}|K_{h,i}^\theta(x) - K_{h,i}^{\theta_{n_k}}(x_{n_k})| &\le \max_{\substack{1\le k\le n^{2p}\\ x\in\mathcal{D}}}\sup_{(\theta,x)\in B_{n_k}} ch^{-2}(|\theta - \theta_{n_k}| + |x - x_{n_k}|) \\
&\le c(nh)^{-1/2}.
\end{aligned}$$

By the strong law of large numbers for dependent observations (see, e.g. Rio, 1995), we have

$$\max_{1\le k\le n^{2p}} \sup_{(\theta,x)\in B_{n_k}} |R_{n,k,2}| \le c(nh)^{-1/2}\frac{1}{n}\sum_{i=1}^{n}|\varphi_i| = O(\delta_n). \tag{6.2}$$

More clearly, we write h as h_n. Let $T_\ell = \{\ell/(h_\ell \log(\ell))\}^\kappa$, where $\kappa = 1/(2r-2)$. Let $\varphi^o_{i,\ell} = \varphi_i I\{|\varphi_i| \ge T_\ell\}$ and $\varphi^I_{i,\ell} = \varphi_i - \varphi^o_{i,\ell}$. We have

$$R_{n,k,1} = \frac{1}{n}\sum_{i=1}^{n}\Big[K^{\theta}_{h,i}(x)\varphi^o_i - E\{K^{\theta}_{h,i}(x)\varphi^o_i\}\Big] + \frac{1}{n}\sum_{i=1}^{n}\xi_{n_k,i}, \tag{6.3}$$

where $\xi_{n_k,i} = K^{\theta_{n_k}}_{h,i}(x_{n_k})\varphi^I_i - E\{K^{\theta_{n_k}}_{h,i}(x_{n_k})\varphi^I_i\}$.

It is easy to check that

$$\sum_{\ell=1}^{\infty}(\ell/h_\ell)^{-1/2}E|\varphi^o_{\ell,\ell}| \le \sum_{\ell=1}^{\infty}(\ell/h_\ell)^{-1/2}T_\ell^{-r+1}E|\varphi_\ell|^r < \infty.$$

Therefore (cf. Rao, 1973, p.111)

$$\sum_{\ell=1}^{\infty}(\ell/h_\ell)^{-1/2}|\varphi^o_{\ell,\ell}| < \infty$$

almost surely. By the Kronecker lemma, we have

$$\frac{1}{n}\sum_{\ell=1}^{n}E|\varphi^o_{\ell,\ell}| = O\{(n/h)^{-1/2}\}, \qquad \frac{1}{n}\sum_{\ell=1}^{n}|\varphi^o_{\ell,\ell}| = O\{(n/h)^{-1/2}\}.$$

Note that $|\varphi^o_{\ell,n}| \le |\varphi^o_{\ell,\ell}|$ for all $\ell \le n$, and $|K^{\theta_{n_k}}_{h,i}(x)| < ch^{-1}$ by (C5). We have

$$\max_{1\le k\le n^{2p}}\frac{1}{n}\sum_{i=1}^{n}E|K^{\theta_{n_k}}_{h,i}(x)\varphi^o_{i,n}| = O\{(nh)^{-1/2}\}, \tag{6.4}$$

$$\max_{1\le k\le n^{2p}}\frac{1}{n}\sum_{i=1}^{n}|K^{\theta_{n_k}}_{h,i}(x)\varphi^o_{i,n}| = O\{(nh)^{-1/2}\}. \tag{6.5}$$

Next, we shall show

$$\max_{1\le k\le n^{2p}}\mathrm{Var}(\sum_{i=1}^{n}\xi_{n_k,i}) \le c_1 n/h. \tag{6.6}$$

By stationarity in (C1), we have

$$\mathrm{Var}(\sum_{i=1}^{n}\xi_{n_k,i}) = n\mathrm{Var}(\xi_{n_k,i}) + 2\sum_{i=2}^{n}(n-i)\mathrm{Cov}(\xi_{n_k,1},\xi_{n_k,i}). \tag{6.7}$$

Let $\tilde{\varphi}_{\theta_{n_k}}(u) = E(|\varphi(X,Z,y)|^\ell \Big| \theta_{n_k}^T X = u)$ and $\tilde{\varphi}_{\theta_{n_k}}(u,v|i) = E(|\varphi_1\varphi_i| \Big| \theta_{n_k}^T X_1 = u, \theta_{n_k}^T X_i = v)$. By the conditions about φ in Lemma A.1 and assumption (C2), we have

$$\begin{aligned}
L(\ell) &\stackrel{def}{=} E\{(K_{h,i}^{\theta_{n_k}}(x_{n_k}))^\ell |\varphi_i|^\ell\} \\
&\le E\{(K_{h,i}^{\theta_{n_k}}(x_{n_k}))^\ell E(|\varphi_i|^\ell \Big| \theta_{n_k}^T X_i)\} \\
&= h^{-\ell} \int (K_h(u - \theta_{n_k}^T x_{n_k}))^\ell \tilde{\varphi}_{\theta_{n_k}}(u) f_{\theta_{n_k}^T X}(u) du \\
&= h^{-\ell+1} \int (K(u))^\ell \tilde{\varphi}_{\theta_{n_k}}(\theta_{n_k}^T x_{n_k} + hu) f_{\theta_{n_k}^T X}(\theta_{n_k}^T x_{n_k} + hu) du \\
&\le ch^{-\ell+1}, \qquad 0 \le \ell \le r, \\
M(i) &\stackrel{def}{=} E\left\{K_{h,1}^{\theta_{n_k}}(x_{n_k}) K_{h,i}^{\theta_{n_k}}(x_{n_k}) |\varphi_1\varphi_i|\right\} \\
&\le E\left\{K_{h,1}^{\theta_{n_k}}(x_{n_k}) K_{h,i}^{\theta_{n_k}}(x_{n_k}) E\left(|\varphi_1\varphi_i| \Big| \theta_{n_k}^T X_1, \theta_{n_k}^T X_i\right)\right\} \\
&= h^{-2} \int K\{(u - \theta_{n_k}^T x_{n_k})/h\} K\{(v - \theta_{n_k}^T x_{n_k})/h\} \tilde{\varphi}_{\theta_{n_k}}(u,v|i) f_{\theta_{n_k}^T X_1, \theta_{n_k}^T X_i}(u,v) du dv \\
&= \int K(u)K(v) \tilde{\varphi}_{\theta_{n_k}}(\theta_{n_k}^T x_{n_k} + hu, \theta_{n_k}^T x_{n_k} + hv|i) \\
&\qquad \times f_{\theta_{n_k}^T X_1, \theta_{n_k}^T X_i}(\theta_{n_k}^T x_{n_k} + hu, \theta_{n_k}^T x_{n_k} + hv) du dv \\
&\le c \int K(u)K(v) \tilde{\varphi}_{\theta_{n_k}}(\theta_{n_k}^T x_{n_k} + hu, \theta_{n_k}^T x_{n_k} + hv|i) du dv \le c \qquad i = 2, 3, \cdots,
\end{aligned}$$

where $f_{\theta_{n_k}^T X}$ and $f_{\theta_{n_k}^T X_1, \theta_{n_k}^T X_i}$ are the density functions of $\theta_{n_k}^T X$ and $(\theta_{n_k}^T X_1, \theta_{n_k}^T X_i)$ respectively. Therefore

$$\mathrm{Var}(\xi_{n_k,i}) \le L(2) \le c/h. \tag{6.8}$$

By the Davydov's lemma (Hall and Heyde, 1980, Corollary 2),

$$\begin{aligned}
|\mathrm{Cov}(\xi_{n_k,1}, \xi_{n_k,i})| &\le 8\{\alpha(i-1)\}^{1-2/r} (E|\xi_{n_k,1}|^r)^{2/r} \\
&\le 8\{\alpha(i-1)\}^{1-2/r} \{L(r)\}^{2/r} \\
&\le ch^{-2+2/r} \{\alpha(i-1)\}^{1-2/r}.
\end{aligned} \tag{6.9}$$

Let $N_1 = INT(h^{(-1+2/r)/(2p)})$, where $INT(v)$ denotes the integer part of v. From (6.7)-(6.9) and assumption (C1), we have

$$\begin{aligned}
\mathrm{Var}(\sum_{i=1}^n \xi_{n_k,i}) &= n\mathrm{Var}(\xi_{n_k,i}) + 2\Big(\sum_{i=2}^{N_1} + \sum_{i=N_1+1}^{n}\Big)(n-i)\mathrm{Cov}(\xi_{n_k,1}, \xi_{n_k,i}) \\
&\le cn/h + 2cn \sum_{i=2}^{N_1} M(i) + 2cnh^{-2+2/r} \sum_{i=N_1+1}^{n} \{\alpha(i-1)\}^{1-2/r}
\end{aligned}$$

$$\begin{aligned} &\leq cn/h + 2cnN_1 + 2cnh^{-2+2/r}N_1^{-2p}\sum_{i=N_1+1}^{n} i^{2p}\{\alpha(i-1)\}^{1-2/r} \\ &\leq cn/h. \end{aligned}$$

Note that c does not depend on k. Therefore (6.6) follows.

Let $N_2 = INT(n^{1/2-1/r}h^{1/2+1/r}(\log n)^{-1/2})$ and $N_3 = INT(n/(2N_2))$. Then $n = 2N_2N_3 + N_0$ and $0 \leq N_0 < 2N_2$. We write

$$W_{n_k}(j) = \sum_{i=(j-1)N_2+1}^{j\cdot N_2} \xi_{n_k,i}, \quad j = 1, \cdots, 2N_2.$$

Then

$$\sum_{i=1}^{n} \xi_{n_k,i} = \sum_{j=1}^{N_3} W_{n_k}(2j-1) + \sum_{j=1}^{N_3} W_{n_k}(2j) + S_{n,0}^T \ , \tag{6.10}$$

where $S_{n,0}^T$ is the residual and has less than $2N_2$ terms. Its contribution is negligible.

For every $\eta > 0$, we use the strong approximation theorem of Bradley (1983) to approximate the random variables $W_{n_k}(1), W_{n_k}(3), \cdots, W_{n_k}(2j-1)$ by independent random variables $W_{n_k}^*(1), W_{n_k}^*(3), \cdots, W_{n_k}^*(2j-1)$ defined as follows. By enlarging the probability space if necessary, introduce a sequence $(U_1, U_2, \cdots)$ of independent uniform $[0,1]$ random variables that are independent of $\{W_{n_k}(1), \cdots, W_{n_k}(2j-1)\}$. Define $W_{n_k}^*(0) = 0, W_{n_k}^*(1) = W_{n_k}(1)$. Then for each $j \geq 2$, there exists a random variable $W_{n_k}^*(2j-1)$ which is a measurable function of $W_{n_k}(1), W_{n_k}(3), \cdots, W_{n_k}(2j-1)$ and U_j such that $W_{n_k}^*(2j-1)$ is independent of $W_{n_k}^*(1), \cdots, W_{n_k}^*(2j-3)$, has the same distributions as $W_{n_k}(2j-1)$ and satisfies

$$P(|W_{n_k}^*(2j-1) - W_{n_k}(2j-1)| > \eta) \leq 18(|W_{n_k}(2j-1)|_\infty/\eta)^{1/2}\alpha(N_2), \tag{6.11}$$

where $|\cdot|_\infty$ is the sup-norm. It follows from the definition of $W_{n_k}^*(2j-1)$ and (6.6) that,

$$EW_{n_k}^*(2j-1) = 0, \quad \max_{k,j} \mathrm{Var}(W_{n_k}^*(2j-1)) \leq c_2 n^{1/2-1/r}h^{-1/2+1/r}(\log n)^{-1/2} \stackrel{def}{=} N_4. \tag{6.12}$$

By the condition in Lemma A.1, we have $h^{-r}(n/\log n)^{-r+2} \to 0$. Hence

$$\begin{aligned} \max_{1\leq k\leq n^{2p}} |\xi_{n_k,i}| &\leq ch^{-1}T_n = c\{n/(h\log n)\}^{1/2}\{h^{-r}(n/\log n)^{-r+2}\}^{\kappa} \\ &\leq c_3\{n/(h\log n)\}^{1/2} \stackrel{def}{=} N_5. \end{aligned} \tag{6.13}$$

Let $N_6 = c_4(nh^{-1}\log n)^{1/2}$. By the Bernstein's inequality, we have from (6.12) and (6.13)

$$\begin{aligned} P(|\sum_{j=1}^{N_3} W_{n_k}^*(2j-1)| > N_6) &\leq \exp\left(\frac{-c_4^2 nh^{-1}\log n}{2(N_3N_4 + N_5N_6)}\right) \\ &\leq \exp\{-c_4^2\log n/(c_2 + 2c_3c_4)\} \\ &\leq c_5 n^{-2p-2}. \end{aligned} \tag{6.14}$$

The last inequality holds if we choose c_4 sufficiently large. By (6.11), if (i) $N_6/N_3 \leq |W^*_{n_k}(2j-1)|_\infty$, we have

$$\begin{aligned}\Pr(|W_{n_k}(2j-1) - W^*_{n_k}(2j-1)| > N_6/N_3) &\leq 18(N_2N_5/(N_6/N_3))^{1/2}\alpha(N_2) \\ &\leq c_6(n/\log n)^{1/2}\alpha(N_2); \end{aligned} \tag{6.15}$$

if (ii) $N_6/N_3 > |W^*_{n_k}(2j-1)|_\infty$, take $\eta = |W^*_{n_k}(2j-1)|_\infty$ in (6.11), we have

$$\Pr(|W_{n_k}(2j-1) - W^*_{n_k}(2j-1)| > \eta) \leq 18\alpha(N_2)\ ,$$

which is smaller than the right hand side of (6.15) as $n \to \infty$. Therefore,

$$\begin{aligned}&\Pr(|\sum_{j=1}^{N_3}\{W_{n_k}(2j-1) - W^*_{n_k}(2j-1)\}| > N_6) \\ &\leq \sum_{j=1}^{N_3}\Pr(|W_{n_k}(2j-1) - W^*_{n_k}(2j-1)| > N_6/N_3) \\ &\leq c_7N_3(n/\log n)^{1/2}\alpha(N_2). \end{aligned} \tag{6.16}$$

From (6.14) and (6.16), we have

$$\begin{aligned}&\Pr(\max_{1\leq k\leq n^{2p}}|\sum_{j=1}^{N_3}W_{n_k}(2j-1)| \geq 2N_6) \\ &\leq \sum_{k=1}^{n^{2p}}\Pr(|\sum_{j=1}^{N_3}W^*_{n_k}(2j-1)| \geq N_6) + \sum_{k=1}^{n^{2p}}\Pr(|\sum_{j=1}^{N_3}|W_{n_k}(2j-1) - W^*_{n_k}(2j-1)| \geq N_6) \\ &\leq n^{2p}\{c_5n^{-2p-2} + c_7N_3(n/\log n)^{1/2}\alpha(N_2)\}.\end{aligned}$$

By (C1), it follows that

$$\sum_{n=1}^{\infty}\Pr(\max_{1\leq k\leq n^{2p}}|\sum_{j=1}^{N_3}W_{n_k}(2j-1)| \geq 2N_6) < \infty.$$

By the Borel-Cantelli lemma, we have

$$\max_{1\leq k\leq n^{2p}}|\sum_{j=1}^{N_3}W_{n_k}(2j-1)| = O(N_6). \tag{6.17}$$

Similarly, we can show

$$\max_{1\leq k\leq n^{2p}}|\sum_{j=1}^{N_3}W_{n_k}(2j)| = O(N_6). \tag{6.18}$$

Combining (6.4), (6.5), (6.10), (6.17), (6.18) and (6.3), we have

$$\max_{1\le k\le n^{2p}} |R_{n,k,1}| = O(\delta_n). \tag{6.19}$$

Therefore, the fourth part of Lemma A.1 follows from (6.1), (6.2) and (6.19).

Note that the key steps in the proof above are the continuity of the related functions and bounded variance in (6.6). To prove the last part of Lemma A.1, it is sufficient to show

$$\sup_{|\theta-\theta_0|\le a_n, x\in\mathcal{D}} E(K^\theta_{h,i}\zeta_i^{k,\ell})^\tau \le ch^{\tau\ell-\tau+1}(a_n^{2\tau}+h^{2\tau}), \quad 2\le\tau\le r. \tag{6.20}$$

Write $\theta_0=\rho_n\theta+\varrho_n\vartheta$, where $\vartheta\perp\theta$, $|\theta|=1$ and $|\vartheta|=1$. It is easy to see that $|\rho_n|<c$ and $|\varrho_n|\sim a_n$ when $|\theta-\theta_0|<a_n$. Let $(\theta,\vartheta,\Gamma)$ be an orthogonal matrix. Let $\tilde f(v,u_1,u_2,\cdots,u_p)$ and $\tilde f(v,u_1,u_2)$ be the density functions of $(\mathbf{z},\theta^TX,\vartheta^TX,\Gamma^TX)$ and $(\mathbf{z},\theta^TX,\vartheta^TX)$ respectively. By (C3), we have

$$\begin{aligned}
E(K^\theta_{h,i}\zeta_i^{k,\ell})^\tau &= \int (K_h(u_1-\theta^Tx))^\tau(u_1-\theta^Tx)^{\tau\ell}v^{\tau k}m^\tau(\rho_nu_1+\varrho_nu_2,\rho_n\theta^Tx+\varrho_n\vartheta^Tx)\\
&\qquad\times\tilde f(v,u_1,u_2\cdots,u_p)dvdu_1du_2\cdots du_p\\
&= h^{\tau\ell-\tau+1}\int (K(v_1))^\tau v_1^{\tau\ell}v^{\tau k}m^\tau(\rho_nv_1h+\rho_n\theta^Tx+\varrho_nu_2,\varrho_n\theta^Tx+\rho_n\vartheta^Tx)\\
&\qquad\times\tilde f(v,\theta^Tx+hv_1,u_2,\cdots,u_p)dvdv_1du_2\cdots du_p\\
&= h^{\tau\ell-\tau+1}\int (K(v_1))^\tau v_1^{\tau\ell}v^{\tau k}m^\tau(\rho_nv_1h+\rho_n\theta^Tx+\varrho_nu_2,\rho_n\theta^Tx+\varrho_n\vartheta^Tx)\\
&\qquad\times\tilde f(v,\theta^Tx+hv_1,u_2)dvdv_1du_2.
\end{aligned}$$

Note that $|m(u,v)|\le c(u-v)^2$. Therefore

$$\begin{aligned}
E(K^\theta_{h,i}\zeta_i^{k,\ell})^\tau &\le ch^{\tau\ell-\tau+1}\int (K(v_1))^\tau v_1^{\tau\ell}v^{\tau k}(\rho_n^{2\tau}v_1^{2\tau}h^{2\tau}+\varrho_n^{2\tau})\tilde f(v,\theta^Tx+hv_1,u_2)dvdv_1du_2\\
&= O\{h^{\tau\ell-\tau+1}(a_n^{2\tau}+h^{2\tau})\}. \qquad\square
\end{aligned}$$

The equations in Lemma A.1 still hold if we replace $|\theta-\theta_0|<a_n$ with $|\theta+\theta_0|<a_n$. The latter is needed for the proof of Theorem 1 in the case $\tilde\theta^T\theta_0<0$.

Lemma A.2. *Let φ_i be defined in Lemma A.1 and $f(x,z,y)$ be the density function of (X,Z,y). If (C1) and (C5) hold, then*

$$\sup_{\theta\in\Theta}\Big|\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\Big\{K^\theta_{h,i}(X_j)\varphi_j-\int K^\theta_{h,i}(x)\varphi(x,z,y)f(x,z,y)dxdzdy\Big\}\varepsilon_i\Big| = O(\delta_n^2).$$

Proof. Let $\Delta_n(\theta)$ be the value in the absolute symbols. By the continuity of $K^\theta_{h,i}$ in θ, there are $n_1<cn^{2p}$ points $\theta_{n,1},\cdots,\theta_{n,n_1}$ in Θ such that $\cup_{k=1}^{n_1}\{\theta:|\theta-\theta_{n,k}|<h^2\delta_n^2\}\supset\Theta$ and

$$\max_{1\le k\le n_1}\sup_{|\theta-\theta_{n,k}|<h^2\delta_n^2}\Big|\Delta_n(\theta)-\Delta_n(\theta_{n,k})\Big| = O(\delta_n^2). \tag{6.21}$$

The Fourier transform $\phi(s) = \int \exp(\mathrm{i}sv)K(v)dv$ will be used in the following, where i is the imaginary unit. Thus $K(v) = \int \exp(-\mathrm{i}sv)\phi(s)ds$. We have

$$\begin{aligned}\Delta_n(\theta_{n,k}) &= \frac{1}{n^2}h^{-1}\sum_{j=1}^n\sum_{i=1}^n\int\Big[\exp\{-\mathrm{i}s\theta_{n,k}^TX_{ij}/h\}\varphi_j \\ &\qquad - \int\exp\{-\mathrm{i}s\theta_{n,k}^TX_{i0}/h\}\varphi(x,z,y)f(x,z,y)dxdzdy\Big]\phi(s)ds\varepsilon_i \\ &= h^{-1}\int\frac{1}{n}\sum_{i=1}^n\exp(-\mathrm{i}s\theta_{n,k}^TX_i/h)\varepsilon_i\cdot\frac{1}{n}\sum_{j=1}^n\Big[\exp(\mathrm{i}s\theta_{n,k}^TX_j/h)\varphi_j \\ &\qquad - \int\exp(\mathrm{i}s\theta_{n,k}^Tx/h)\varphi(x,z,y)f(x,z,y)dxdzdy\Big]\phi(s)ds.\end{aligned}$$

Following the same steps leading to (6.19), we have

$$\max_{1\le k\le n_1}|\frac{1}{n}\sum_{i=1}^n\exp(-\mathrm{i}s\theta_{n,k}^TX_i/h)\varepsilon_i| \le c_8\delta_{0n},$$

$$\max_{1\le k\le n_1}\Big|\frac{1}{n}\sum_{j=1}^n\Big[\exp(\mathrm{i}s\theta_{n,k}^TX_j/h)\varphi_j - \int\exp(\mathrm{i}s\theta_{n,k}^Tx/h)\varphi(x,z,y)f(x,z,y)dxdzdy\Big]\Big| \le c_9\delta_{0n}$$

almost surely, where c_8 and c_9 are constants which do not depend on s. Hence

$$\max_{1\le k\le n_1}\Big|\Delta_n(\theta_{n,k})\Big| \le h^{-1}\int c_8\delta_{0n}c_9\delta_{0n}|\phi(s)|ds = O(h^{-1}\delta_{0n}^2) = O(\delta_n^2). \tag{6.22}$$

Note that

$$\sup_{\theta\in\Theta}|\Delta_n(\theta)| \le \max_{1\le k\le n_1}\Big|\Delta_n(\theta_{n,k})\Big| + \max_{1\le k\le n_1}\sup_{|\theta-\theta_{n,k}|<h^2\delta_n^2}\Big|\Delta_n(\theta)-\Delta_n(\theta_{n,k})\Big|. \tag{6.23}$$

Therefore, the second part of Lemma A.2 follows from (6.21), (6.22) and (6.23). □

For easy of exposition, we abuse $\mathcal{D}$ as the positive support of the $f(x)$. Let $d(x,\mathcal{D}^c) = \min_{x'\in\mathbb{R}^p-\mathcal{D}}|x-x'|$ and define bounded functions $J_0(x)$, $J_\theta(v)$ such that $J_0(x) = 0$ if $d(x,\mathbb{R}^p-\mathcal{D}) > a_0b$ and $J_\theta(\theta^Tx) = 0$ if $d(\theta^Tx,\theta^T(\mathbb{R}^p-\mathcal{D})) > b_0h$. By the definition, we have

$$\frac{1}{n}\sum_{j=1}^n J_0(X_j) = O(b), \qquad \frac{1}{n}\sum_{j=1}^n J_\theta(X_j) = O(h). \tag{6.24}$$

To cope with the boundary points, we give the following nonuniform rate of convergency.

Lemma A.3. *Suppose assumptions (C3) and (C5) hold. Then*

$$\begin{aligned}EH_b(X-x)\{\theta^T(X-x)/b\}^k\{\vartheta^T(X-x)/b\}^\ell &= v_{k,\ell}^{\theta,\vartheta}f(x) + J_0(x) + O(h), \\ EK_h(\theta^T(X-x))\{\theta^T(X-x)/h\}^\ell &= \tau_\ell f_\theta(\theta^Tx) + J_\theta(x) + O(h),\end{aligned}$$

uniformly for $\theta,\vartheta\in\Theta$ and $x\in\mathcal{D}$, where $v_{k,\ell}^{\theta,\vartheta} = \int_{\mathbb{R}^p}H(U)(\theta^TU)^k(\vartheta^TU)^\ell dU$ and $\tau_\ell = \int K(u)u^\ell du$.

Proof. We here only give the details for the first part. If $d(x, \mathcal{D}^c) > a_0 b$, we define $J_0(x) = 0$. From (C5), we have

$$\begin{aligned}&\int_{\mathcal{D}} H_b(U-x)\{\theta^T(U-x)/b\}^k\{\vartheta^T(U-x)/b\}^\ell f(U)dU\\ &= \int_{\mathbb{R}^p} H(U)\{\theta^T U\}^k\{\theta^T U\}^\ell f(x+hU)dU = v_{k,\ell}^{\theta,\vartheta} f(x) + O(h).\end{aligned}$$

If $d(x, \mathcal{D}^c) < a_0 b$, we have by (C3)

$$\begin{aligned}J_0(x) &\stackrel{def}{=} \int_{\mathcal{D}} H_b(U-x)\{\theta^T(U-x)/b\}^k\{\vartheta^T(U-x)/b\}^\ell f(U)dU\\ &\le \int_{\mathbb{R}^p} H(U)\{\theta^T U\}^k\{\vartheta^T U\}^\ell f(x+hU)dU = O(1).\end{aligned}$$

Therefore, the first part of Lemma A.3 follows. □

In the following context, we abbreviate L for any function $L(x)$, and L_θ or $L_\theta(x)$ for any function $L_\theta(\theta^T x)$. Let ν_θ and μ_θ be defined as in section 2, and

$$\nu = E(Z|X=x), \quad \pi = E(ZZ^T|X=x), \quad \pi_\theta = E(ZZ^T|\theta^T X = \theta^T x),$$
$$\tilde{\Sigma}_\theta = E(XX^T|\theta^T X = \theta^T x) - \mu_\theta x^T - x\mu_\theta^T + xx^T.$$

Let

$$\begin{aligned}\varsigma_0 &= \frac{1}{n}\sum_{i=1}^n H_{b,i}, \quad S_1 = \frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}, \quad S_2 = \frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}X_{i0}^T,\\ T_1 &= \frac{1}{n}\sum_{i=1}^n H_{b,i}Z_i, \quad T_2 = \frac{1}{n}\sum_{i=1}^n H_{b,i}Z_iZ_i^T, \quad C_2 = \frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}Z_i^T,\\ E_1 &= \frac{1}{n}\sum_{i=1}^n H_{b,i}Z_iy_i, \quad D_1 = \frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}y_i, \quad W_n = \varsigma_0 S_2 - S_1S_1^T\end{aligned}$$

and

$$\bar{w}_{a,i}^\theta(x) = \{\theta^T S_2\theta\}H_{b,i} - \theta^T S_1 H_{b,i}\theta^T X_{i0}, \quad \bar{w}_{d,i}^\theta(x) = \varsigma_0 H_{b,i}\theta^T X_{i0} - \theta^T S_1 H_{b,i}.$$

Based on (2.4), we can obtain initial estimators of θ_0 and β_0 as follows. Choose a vector θ with norm 1 and any vector β. Let $\bar{w}_j^\theta = \theta^T W_n(X_j)\theta$ and calculate

$$\bar{a}_j^\theta = \{\bar{w}_j^\theta\}^{-1}\sum_{i=1}^n \bar{w}_{a,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad \bar{d}_j^\theta = \{\bar{w}_j^\theta\}^{-1}\sum_{i=1}^n \bar{w}_{d,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad (6.25)$$

$$\begin{pmatrix}\bar{\beta}\\ \bar{\theta}\end{pmatrix} = \{\bar{D}_n^\theta\}^- \sum_{j=1}^n I_n(X_j)\begin{pmatrix}E_1(X_j) - \bar{a}_j^\theta T_1(X_j)\\ \bar{d}_j^\theta D_1(X_j) - \bar{a}_j^\theta\bar{d}_j^\theta S_1(X_j)\end{pmatrix}/\varsigma_0(X_j), \quad \bar{\theta} := \bar{\theta}/|\bar{\theta}|, \quad (6.26)$$

where

$$\bar{D}_n^\theta = \sum_{j=1}^n I_n(X_j) \begin{pmatrix} T_2(X_j) & \bar{d}_j^\theta C_2(X_j) \\ \bar{d}_j^\theta C_2^T(X_j) & (\bar{d}_j^\theta)^2 S_2(X_j) \end{pmatrix} / \varsigma_0(X_j),$$

and A^- denotes the Moore-Penrose inverse of matrix A. Repeat the calculations in (6.25) and (6.26) with (θ, β) replaced by $(\bar{\theta}, \bar{\beta})$ until convergence. Denote the final value by $(\tilde{\beta},\ \tilde{\theta})$. Next, we shall improve the efficiency of the estimators using a univariate kernel. Let

$$\varsigma_k^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta \{\theta^T X_{i0}\}^k, \quad k = 0, 1, 2, 3,$$

$$w_{a,i}^\theta = \varsigma_2^\theta K_{h,i}^\theta - \varsigma_1^\theta K_{h,i}^\theta \theta^T X_{i0}, \quad w_{d,i}^\theta = \varsigma_0^\theta K_{h,i}^\theta \theta^T X_{i0} - \varsigma_1^\theta K_{h,i}^\theta,$$

$$w^\theta = \frac{1}{n}\sum_{i=1}^n w_{a,i}^\theta, \quad S_1^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta X_{i0}, \quad S_2^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta X_{i0} X_{i0}^T,$$

$$T_1^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta Z_i, \quad E_1^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta Z_i y_i, \quad D_1^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta X_{i0} y_i,$$

$$T_2^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta Z_i Z_i^T, \quad C_2^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta \theta^T X_{i0} Z_i^T,$$

$$S_{1,1}^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta \{\theta^T X_{i0}\} X_{i0}, \quad S_{2,1}^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta \{\theta^T X_{i0}\}^2 X_{i0},$$

$$S_{1,2}^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta \{\theta^T X_{i0}\} X_{i0} X_{i0}^T, \quad S_3^\theta = \frac{1}{n}\sum_{i=1}^n K_{h,i}^\theta X_{i0} \{(\theta - \theta_0)^T X_{i0}\}^2.$$

Based on (2.4), we improve the estimators $\tilde{\theta}$ and $\tilde{\beta}$ as follows. Let $w_j^\theta = w^\theta(X_j)$. Starting with $(\theta, \beta) = (\tilde{\theta}, \tilde{\beta})$, calculate

$$\tilde{a}_j^\theta = (w_j^\theta)^{-1} \sum_{i=1}^n w_{a,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \quad \tilde{d}_j^\theta = (w_j^\theta)^{-1} \sum_{i=1}^n w_{d,i}^\theta(X_j)\{y_i - \beta^T Z_i\}, \tag{6.27}$$

$$\begin{pmatrix} \tilde{\beta} \\ \tilde{\theta} \end{pmatrix} = (\tilde{D}_n^\theta)^- \sum_{j=1}^n I_n(X_j) \begin{pmatrix} E_1^\theta(X_j) - \tilde{a}_j^\theta T_1^\theta(X_j) \\ \tilde{d}_j^\theta D_1^\theta(X_j) - \tilde{a}_j^\theta \tilde{d}_j^\theta S_1^\theta(X_j) \end{pmatrix} / \varsigma_0^\theta(X_j), \quad \tilde{\theta} := \tilde{\theta}/|\tilde{\theta}|, \tag{6.28}$$

where

$$\tilde{D}_n^\theta = \sum_{j=1}^n I_n(X_j) \begin{pmatrix} T_2^\theta(X_j) & \tilde{d}_j^\theta C_2^\theta(X_j) \\ \tilde{d}_j^\theta \{C_2^\theta(X_j)\}^T & (\tilde{d}_j^\theta)^2 S_2^\theta(X_j) \end{pmatrix} / \varsigma_0^\theta(X_j).$$

Repeat the procedure (6.27) and (6.28) with (θ, β) replaced by $(\tilde{\theta}, \tilde{\beta})$ until convergence. Denote the final value by $(\hat{\beta},\ \hat{\theta})$.

Let $\bar{\Delta}_i(x) = y_i - \bar{a} - \beta_0^T Z_i - \bar{d} X_{i0}^T \theta_0$ and $\tilde{\Delta}_i^\theta(x) = y_i - \tilde{a} - \beta_0^T Z_i - \tilde{d} X_{i0}^T \theta_0$. We have

$$\begin{pmatrix} \bar{\beta} \\ \bar{\theta} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \theta_0 \end{pmatrix} + \bar{D}_n^-(\theta) \sum_{j=1}^n I_n(X_j) \sum_{i=1}^n H_{b,i}(X_j) \begin{pmatrix} Z_i \\ X_{ij} \bar{d}_j \end{pmatrix} \bar{\Delta}_i(X_j)/\varsigma_0(X_j), \tag{6.29}$$

$$\begin{pmatrix} \tilde{\beta} \\ \tilde{\theta} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \theta_0 \end{pmatrix} + \tilde{D}_n^-(\theta) \sum_{j=1}^n I_n(X_j) \sum_{i=1}^n K_{h,i}^\theta(\theta^T X_j) \begin{pmatrix} Z_i \\ X_{ij}\tilde{d}_j \end{pmatrix} \tilde{\Delta}_i^\theta(X_j)/\varsigma_0^\theta(\theta^T X_j). \quad (6.30)$$

By Lemmas A.1 and A.3, we have

$$\begin{gathered} \varsigma_0 = f(x) + O(J_0 + \tau_{pn}), \quad S_1 = O(bJ_0 + b\tau_{pn}), \\ S_2 = f(x) I_{p\times p} b^2 + O(b^2 J_0 + b^2\tau_{pn}), \quad T_1 = f(x)\nu(x) + O(J_0 + \tau_{pn}), \\ T_2 = f(x)\pi(x) + O(J_0 + \tau_{pn}), \quad \frac{1}{n}\sum_{i=1}^n H_{b,i} Z_i \varepsilon_i = O(\delta_{pn}), \\ \frac{1}{n}\sum_{i=1}^n H_{b,i}\varepsilon_i = O(\delta_{pn}), \quad \frac{1}{n}\sum_{i=1}^n H_{b,i} X_{i0}\{\theta^T X_{i0}\}^k \varepsilon_i = O(b^{k+1}\delta_{pn}), \\ \frac{1}{n}\sum_{i=1}^n H_{b,i}|X_{i0}|^k = O(b^k), \quad C_2 = O(bJ_0 + b^2 + b\delta_n), \end{gathered} \quad (6.31)$$

and

$$\begin{gathered} \varsigma_0^\theta = f_\theta + O(J_\theta + \tau_n), \quad \varsigma_1^\theta = O(hJ_\theta + h^2 + h\delta_n), \quad \varsigma_2^\theta = f_\theta h^2 + O(h^2 J_\theta + h^2\tau_n), \\ \varsigma_3^\theta = O(h^4 + b^3 J_\theta + h^3\delta_n), \quad S_1^\theta = f_\theta\{\mu_\theta - x\} + O(J_\theta + \tau_n), \quad S_2^\theta = \tilde{\Sigma}_\theta f_\theta + O(J_\theta + \tau_n), \\ w^\theta = f_\theta^2 h^2 + O(h^2 J_\theta + h^2\tau_n), \quad T_1^\theta = f_\theta\nu_\theta + O(J_\theta + \tau_n), \quad T_2^\theta = f_\theta\pi_\theta + O(J_\theta + \tau_n), \\ C_2^\theta = O(hJ_\theta + h^2 + h\tau_n), \quad S_{1,1}^\theta = O(hJ_\theta + h^2 + h\tau_n), \quad S_{1,2}^\theta = O(hJ_\theta + h^2 + h\tau_n), \\ S_{2,1}^\theta = f_\theta\{\mu_\theta - x\}h^2 + O(h^2 J_\theta + h^2\tau_n), \quad S_3^\theta = O(\delta_\theta^2). \end{gathered} \quad (6.32)$$

Let $\bar{a}, \bar{d}, \tilde{a}$ and $\tilde{d}$ be respectively the values of $\bar{a}_j, \bar{d}_j, \tilde{a}_j$ and $\tilde{d}_j$ with X_j replaced by x. For simplicity, we further assume that $f(x) > c_0$ and $f_\theta(\theta^T x) > c_0$ for all $x \in \mathcal{D}$ (otherwise, we only need to change $\mathcal{D}$ to $\{x : f(x) > c_0\}$ or $\{x : f_\theta(\theta^T x) > c_0\}$ in the proofs). Thus, $I_n(X_j) \equiv 1$ when n is sufficiently large.

Lemma A.4. *Let $\beta_d = \beta_0 - \beta$ and $\theta_d = \theta_0 - \theta$. Suppose assumptions (C1)-(C5) hold. We have*

$$\begin{aligned} \bar{a} &= g(\theta_0^T x) + \nu^T\beta_d + O(J_0 + b + \delta_{pn}), \\ \bar{d} &= \theta^T\theta_0 g'(\theta_0^T x) + O\{(1 + b^{-1}J_0)\delta_\beta + b^{-1}\delta_{pn} + b\}, \\ \tilde{a} &= g(\theta_0^T x) + g'(\theta_0^T x)\{\mu_\theta - x\}^T\theta_d + \nu_\theta^T\beta_d + \tfrac{1}{2}g''(\theta_0^T x)h^2 + R_{n,3} \\ &\quad + O(\delta_\theta^2 + J_\theta\delta_\gamma + \tau_n\delta_\gamma + h\tau_n), \\ \tilde{d} &= g'(\theta_0^T x) + h^{-1}R_{n,4} + O\{\delta_\theta^2 + (h^{-1}J_\theta + 1 + h^{-1}\delta_n)\delta_\gamma + \tau_n\} \end{aligned}$$

uniformly for $x \in \mathcal{D}$ and $\theta \in \Theta$, where $R_{n,3} = \{nf_\theta\}^{-1}\sum_{i=1}^n K_{h,i}^\theta\varepsilon_i$ and $R_{n,4} = \{nhf_\theta\}^{-1}$ $\sum_{i=1}^n K_{h,i}^\theta\theta^T X_{i0}\varepsilon_i$.

Proof. By assumption (C4), we have the following Taylor expansion

$$y_i = \beta_0^T Z_i + g(\theta_0^T x) + g'(\theta_0^T x)\theta_0^T X_{i0} + \frac{1}{2} g''(\theta_0^T x)\{\theta_0^T X_{i0}\}^2 + m(\theta_0^T X_i, \theta_0^T x) + \varepsilon_i, \quad (6.33)$$

where $m(\theta_0^T X_i, \theta_0^T x)$ is defined as in Lemma A.1. Because $\theta^T\theta = 1$, we have by the set of equations in (6.31),

$$W_n = f^2 I_{p\times p} b^2 + O(b^2 J_0 + b^2\tau_{pn}), \quad \frac{1}{n}\sum_{i=1}^n \bar{w}_{a,i}^\theta = \theta^T W_n \theta,$$
$$\frac{1}{n}\sum_{i=1}^n \bar{w}_{a,i}^\theta X_{i0}^T\theta_0 = O(b^3), \quad \frac{1}{n}\sum_{i=1}^n \bar{w}_{a,i}^\theta \{X_{i0}^T\theta_0\}^2 = O(b^3), \quad \frac{1}{n}\sum_{i=1}^n \bar{w}_{a,i}^\theta \varepsilon_i = O(b^2\delta_{pn}),$$
$$\frac{1}{n}\sum_{i=1}^n \bar{w}_{a,i}^\theta Z_i^T = \theta^T S_2\theta T_1^T - \theta^T S_1\theta^T C_2 = f^2\nu^T b^2 + O(b^2 J_0 + b^2\tau_{pn}). \quad (6.34)$$

Combining the equations in (6.34), (6.33) and (6.25), we have the first part of Lemma A.4.

By the definition of $\bar{w}_{d,i}^\theta$ and the set of equations in (6.31), we have

$$\frac{1}{n}\sum_{i=1}^n \bar{w}_{d,i}^\theta = 0, \quad \frac{1}{n}\sum_{i=1}^n \bar{w}_{d,i}^\theta \{X_{i0}^T\theta_0\}^2 = O(b^3 J_0 + b^3\delta_n + b^4),$$
$$\frac{1}{n}\sum_{i=1}^n \bar{w}_{d,i}^\theta X_{i0}^T\theta_0 = \theta^T W_n\theta_0 = \theta^T\theta_0 f^2 b^2 + O(b^2 J_0 + b^2\tau_{pn}),$$
$$\frac{1}{n}\sum_{i=1}^n \bar{w}_{d,i}^\theta Z_i^T = \theta^T S_1 T_1^T - \varsigma_0\theta^T C_2 = O(b^2 + bJ_0 + b\delta_{pn}),$$
$$\frac{1}{n}\sum_{i=1}^n \bar{w}_{d,i}^\theta \varepsilon_i = \theta^T S_1 \frac{1}{n}\sum_{i=1}^n H_{b,i}\varepsilon_i - \varsigma_0 \frac{1}{n}\sum_{i=1}^n H_{b,i}\theta^T X_{i0}\varepsilon_i = O(b\delta_{pn}). \quad (6.35)$$

Combining the equations in (6.35), (6.33) and (6.25), we have the second part of Lemma A.4.

Write $\theta_0 = \theta_d + \theta$. We have by the set of equations in (6.32)

$$\frac{1}{n}\sum_{i=1}^n w_{a,i}^\theta = w^\theta = f_\theta^2 h^2 + O(h^2 J_\theta + h^2\tau_n),$$
$$\frac{1}{n}\sum_{i=1}^n w_{a,i}^\theta X_{i0}^T\theta_0 = \varsigma_2^\theta\theta_d^T S_1^\theta - \varsigma_1^\theta\theta_d^T S_{1,1}^\theta = f_\theta^2\{\mu_\theta - x\}^T\theta_d h^2 + O\{h^2(J_\theta + \tau_n)\delta_\theta\},$$
$$\frac{1}{n}\sum_{i=1}^n w_{a,i}^\theta \{X_{i0}^T\theta_0\}^2 = \varsigma_2^\theta\Big\{\theta_d^T S_2^\theta\theta_d + 2\theta_d^T S_{1,1}^\theta + \varsigma_2^\theta\Big\} - \varsigma_1^\theta\Big\{\theta_d^T S_{2,1}^\theta\theta_d + 2\theta_d^T S_{1,2}^\theta + \varsigma_3^\theta\Big\}$$
$$= f_\theta^2 h^4 + O\{J_\theta h^4 + h^5 + h^2\delta_\theta^2 + h^2(h^2 + J_\theta)\delta_\theta\},$$
$$\frac{1}{n}\sum_{i=1}^n w_{a,i}^\theta Z_i = \varsigma_2^\theta T_1^\theta - \varsigma_1^\theta T_{1,1}^\theta = f_\theta^2\nu_\theta h^2 + O\{h^2(J_\theta + \tau_n)\},$$
$$\frac{1}{n}\sum_{i=1}^n w_{a,i}^\theta m(\theta_0^T X_i, \theta_0^T x) = O\{(h^2 + \delta_\theta^2)(h + J_\theta) + h^2\delta_n(\delta_\theta^2 + h^2)\},$$

$$\frac{1}{n}\sum_{i=1}^{n} w_{a,i}^{\theta}\varepsilon_i = \varsigma_2^{\theta} f_\theta R_{3,n} - \varsigma_1^{\theta} f_\theta h R_{4,n} = f_\theta^2 h^2 R_{3,n} + O\{h^2(h+J_\theta)\delta_n\}. \tag{6.36}$$

Therefore, the third part of Lemma A.4 follows from (6.27), (6.33) and the set of equations in (6.36).

Similarly, we have by the set of equations in (6.32) and Lemma A.1 and A.3, $n^{-1}\sum_{i=1}^{n} w_{d,i}^{\theta} = 0$ and

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} w_{d,i}^{\theta}\{X_{i0}^T\theta_0\} = w^{\theta} + \{\varsigma_0^{\theta} S_{1,1}^{\theta} - \varsigma_1^{\theta} S_1^{\theta}\}^T\theta_d = w^{\theta} + O\{(hJ_\theta + h^2)\delta_\theta\},\\
&\frac{1}{n}\sum_{i=1}^{n} w_{d,i}^{\theta}\{X_{i0}^T\theta_0\}^2 = \varsigma_0^{\theta}\Big\{\varsigma_3^{\theta} + 2\theta_d^T S_{1,2}^{\theta} + \theta_d^T S_{1,2}^{\theta}\theta_d\Big\} - \varsigma_1^{\theta}\Big\{\varsigma_2^{\theta} + 2\theta_d^T S_{1,1}^{\theta} + \theta_d^T S_2^{\theta}\theta_d\Big\}\\
&\qquad\qquad = O(h^4 + h^3 J_\theta + hJ_\theta\delta_\theta + h^2\delta_\theta),\\
&\frac{1}{n}\sum_{i=1}^{n} w_{d,i}^{\theta} m(\theta_0^T X_i, \theta_0^T x) = O\{h(h^2+\delta_\theta^2)(h+J_\theta) + h\delta_n(\delta_\theta^2 + h^2)\},\\
&\frac{1}{n}\sum_{i=1}^{n} w_{d,i}^{\theta} Z_i = \varsigma_0^{\theta} C_2^{\theta} - \varsigma_1^{\theta} T_1^{\theta} = O(h^2 + h\delta_n + hJ_\theta),\\
&\frac{1}{n}\sum_{i=1}^{n} w_{d,i}^{\theta}\varepsilon_i = \varsigma_0^{\theta} h f_\theta R_{4,n} - \varsigma_1^{\theta} f_\theta R_{3,n} = h f_\theta^2 R_{4,n} + O\{(h^2 + hJ_\theta)\delta_n\}.
\end{aligned} \tag{6.37}$$

The last part of Lemma A.4 follows from the equations in (6.37), (6.33) and (6.27). □

To prove Theorem 1 for the case that $\tilde{\theta}^T\theta_0 < 0$, we need to change $\theta_d = \theta_0 - \theta$ and $g'(\theta_0^T x)$ in Lemma A.4 to $\theta_d = -\theta_0 - \theta$ and $-g'(\theta_0^T x)$ respectively.

Lemma A.5. *Suppose assumptions (C1)-(C5) hold. We have*

$$\frac{1}{n}\bar{D}_n(\theta) = \begin{pmatrix} E(ZZ^T) + O(b+\delta_{pn}) & O(b^2 + b\delta_{pn}) \\ O(b^2 + b\delta_{pn}) & (\theta^T\theta_0)^2 E\{g'(\theta_0^T X)\}^2 I_{p\times p} b^2 + O(b\delta_{pn} + b^2\delta_\beta) \end{pmatrix},$$

and

$$\frac{1}{n}\tilde{D}_n(\theta) = \begin{pmatrix} E(ZZ^T) & \tilde{C}_{12} \\ \tilde{C}_{12}^T & 2\tilde{W}_0 \end{pmatrix} + O(h^{-1}\delta_n + \delta_\gamma),$$

uniformly for $\theta \in \Theta$, *where* $\tilde{C}_{12} = E\{g'(\theta_0^T X)Z(\mu_{\theta_0}(X) - X)^T\}$ *and* $\tilde{W}_0 = E[\{g'(\theta_0^T X)\}^2\{X - \mu_{\theta_0}(X)\}\{X - \mu_{\theta_0}(X)\}^T]$.

Proof . To prove Lemma A.5, it is sufficient to show that

$$\frac{1}{n}\sum_{j=1}^{n} T_2(X_j)/\varsigma_0(X_j) = E(ZZ^T) + O(b+\delta_{pn}), \quad \frac{1}{n}\sum_{j=1}^{n} \bar{d}_j^{\theta} C_2(X_j)/\varsigma_0(X_j) = O(b^2 + b\delta_{pn}),$$

$$\frac{1}{n}\sum_{j=1}^{n} (\bar{d}_j^{\theta})^2 S_2(X_j)/\varsigma_0(X_j) = (\theta^T\theta_0)^2 E\{g'(\theta_0^T X)\}^2 I_{p\times p} b^2 + O\{b^2(\delta_\beta + b^{-1}\tau_{pn})\},$$

$$\frac{1}{n}\sum_{j=1}^{n} T_2^{\theta}(\theta^T X_j)/\varsigma_0^{\theta}(\theta^T X_j) = E(ZZ^T) + O(h^{-1}\delta_n + \delta_\gamma),$$

$$\frac{1}{n}\sum_{j=1}^{n} \bar{d}_j^{\theta} C_2^{\theta}(\theta^T X_j)/\varsigma_0^{\theta}(\theta^T X_j) = \tilde{C}_{12} + O(h^{-1}\delta_n + \delta_\gamma),$$

$$\frac{1}{n}\sum_{j=1}^{n} (\bar{d}_j^{\theta})^2 S_2^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = 2\tilde{W}_0 + O(h^{-1}\delta_n + \delta_\gamma).$$

Here, we give the details for the last equation. The other equations can be proved similarly. By Lemma A.1 and that $R_{n,4} = O(h^{-1}\delta_n)$, we have

$$\tilde{d} = g'(\theta_0^T x) + O\{h^2 + h^{-1}\delta_n + (1 + h^{-1}J_\theta)\delta_\gamma\}.$$

By (C2) and (C3), $\tilde{\Sigma}_\theta$ has bounded derivative in θ. By the equations in (6.32), (6.24) and the first part of Lemma A,1, we have

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} (\bar{d}_j^{\theta})^2 S_2^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) &= \frac{1}{n}\sum_{j=1}^{n} \{g'(\theta_0^T X_j)\}^2 \tilde{\Sigma}_\theta(X_j) + O(h^{-1}\delta_n + \delta_\gamma) \\
&= \frac{1}{n}\sum_{j=1}^{n} \{g'(\theta_0^T X_j)\}^2 \tilde{\Sigma}_{\theta_0}(X_j) + O(h^{-1}\delta_n + \delta_\gamma) \\
&= 2\tilde{W}_0 + O(\delta_{0n}) + O(h^{-1}\delta_n + \delta_\gamma) \\
&= 2\tilde{W}_0 + O(h^{-1}\delta_n + \delta_\gamma). \qquad \square
\end{aligned}$$

Lemma A.6. *Suppose assumptions (C1)-(C5) hold. Then*

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} H_{b,i}(X_j) Z_i \bar{\Delta}_i(X_j)/\varsigma_0(X_j) = E\{\nu(X)\nu^T(X)\}(\beta - \beta_0) + O(b + \delta_{pn}),$$

$$\frac{1}{n^2}\sum_{j=1}^{n} \bar{d}_j \sum_{i=1}^{n} H_{b,i}(X_j) X_{ij} \bar{\Delta}_i(X_j)/\varsigma_0(X_j) = b^2(\theta^T\theta_0)(1 - \theta^T\theta_0) E\{g'(\theta_0^T X)\}^2\theta_0 + O(b^3 + b\delta_{pn} + b^2\delta_\beta),$$

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) Z_i \tilde{\Delta}_i^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = E\{\nu_\theta(X)\nu_\theta^T(X)\}\beta_d + \frac{1}{n}\sum_{i=1}^{n}\{Z_i - \nu_\theta(X_i)\}\varepsilon_i + O\{(\delta_\theta + h + h^{-1}\delta_n)\delta_\gamma + h\tau_n\},$$

$$\frac{1}{n^2}\sum_{j=1}^{n} \tilde{d}_j \sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij} \tilde{\Delta}_i^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = \tilde{W}_0\theta_d + \frac{1}{n}\sum_{i=1}^{n} g'(\theta_0^T X_i)\{\mu_{\theta_0}(X_j) - X_i\}\varepsilon_i + O\{(\delta_\gamma + h^{-1}\delta_n + h)\delta_\gamma + h\tau_n + h^{-1}\delta_n^2\},$$

uniformly for $\theta \in \Theta$.

Proof. By Lemma A.4 and expansion (6.33), we have

$$\bar{\Delta}_i = \varepsilon_i + (1 - \theta^T\theta_0) g'(\theta_0^T x) X_{i0}^T\theta_0 - \nu^T(\beta_0 - \beta) + Q_{n,i},$$

where $Q_{n,i} = O\Big(\delta_{pn} + J_0\delta_\beta + b^2 + \{(1+b^{-1}J_0)\delta_\beta + b^{-1}\delta_{pn} + b\}|X_{i0}| + |X_{i0}|^2\Big)$. It follows from (6.24), the equations in (6.32) and Lemmas A.1 and A.3 that

$$\frac{1}{n^2}\sum_{j=1}^n\sum_{i=1}^n H_{b,i}(X_j)Z_i\bar{\Delta}_{n,i}(X_j)/\varsigma_0(X_j) = \frac{1}{n}\sum_{j=1}^n \nu(X_j)\nu(X_j)(\beta-\beta_0) + O(b+\delta_{pn}).$$

Therefore, the first part of Lemma A.6 follows from the first part of Lemma A.1 by taking $m_1(\theta, X, Z) = \nu(X)\nu^T(X)$. Note that by Lemmas A.1 and A.3,

$$\bar{d} = (\theta^T\theta_0)g'(\theta_0^T x) + O\{(1+b^{-1}J_0)\delta_\beta + b^{-1}\delta_{pn} + b\}. \tag{6.38}$$

It follows from the equations in (6.32) that

$$\frac{1}{n^2}\sum_{j=1}^n \bar{d}_j\sum_{i=1}^n H_{b,i}(X_j)X_{ij}\varepsilon_i = O(b\delta_{pn}). \tag{6.39}$$

Note that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}(1-\theta^T\theta_0)X_{i0}^T\theta_0 &= b^2 f(1-\theta^T\theta_0)\theta_0 + O(b^2J_0 + b^2\tau_{pn}),\\
\frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}\{X_{i0}^T\theta_0\}^2 &= O(b^3J_0 + b^4 + b^3\delta_{pn}),\\
\frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}\nu^T\beta_d &= O\{(b^2 + bJ_0 + b\delta_{pn})\delta_\beta\},\\
\frac{1}{n}\sum_{i=1}^n H_{b,i}X_{i0}Q_{n,i} &= O\{b^3 + (b^2+bJ_0)\delta_\beta + b\delta_{pn}\}.
\end{aligned}$$

Hence by the foregoing set of equations and (6.38), we have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{j=1}^n \bar{d}_j\sum_{i=1}^n H_{b,i}(X_j)X_{ij}\bar{\Delta}_{n,i}(X_j)/\varsigma_0(X_j)\\
&= b^2\frac{1}{n}\sum_{j=1}^n \bar{d}_j g'(\theta_0^T X_j)(1-\theta_0^T\theta)\theta_0 + O(b^3 + b\delta_{pn} + b^2\delta_\beta)\\
&= b^2\frac{1}{n}\sum_{j=1}^n \{g'(\theta_0^T X_j)\}^2(\theta^T\theta_0)(1-\theta^T\theta_0)\theta_0 + O(b^3 + b\delta_{pn} + b^2\delta_\beta).
\end{aligned}$$

Therefore, the second part of Lemma A.6 follows from the foregoing equation and the first part of Lemma A.1.

By the expansions of $\tilde{a}$ and $\tilde{d}$ in Lemma A.4, we have

$$\begin{aligned}
\tilde{\Delta}_i^\theta &= \{\varepsilon_i - R_{n,3} - X_{i0}^T\theta_0 R_{n,4}\} + \frac{1}{2}\{(X_{i0}^T\theta_0)^2 - h^2\}g''(\theta_0^T x) - g'(\theta_0^T x)\{\mu_\theta - x\}^T\theta_d\\
&\quad -\nu_\theta^T\beta_d + m(\theta_0^T X_i, \theta_0^T x) + O(\delta_\theta^2 + J_\theta\delta_\gamma + \tau_n\delta_\gamma + h\tau_n)\\
&\quad +O\{\delta_\theta^2 + (h + h^{-1}J_\theta + h^{-1}\delta_n)\delta_\gamma + \tau_n\}|\theta_0^T X_{i0}| \stackrel{def}{=} \sum_{k=1}^7 \tilde{\Delta}_{k,i}^\theta.
\end{aligned} \tag{6.40}$$

By the set of equations in (6.24) and (6.32), we have

$$\begin{aligned}\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)Z_i\varepsilon_i/\varsigma_0^{\theta}(\theta^T X_j) &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)f_{\theta}^{-1}(\theta^T X_j)Z_i\varepsilon_i + O(h\delta_n)\\ &= \frac{1}{n}\sum_{i=1}^{n}Z_i\varepsilon_i\{\frac{1}{n}\sum_{j=1}^{n}K_{h,i}^{\theta}(X_j)f_{\theta}^{-1}(\theta^T X_j)\} + O(h\delta_n)\\ &= \frac{1}{n}\sum_{i=1}^{n}Z_i\varepsilon_i + O(h\delta_n). \qquad (6.41)\end{aligned}$$

Note that $R_{n,3} = O(\delta_n)$. It follows from Lemmas A.1 and A.2 that

$$\begin{aligned}&\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)Z_iR_{n,3}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = \frac{1}{n}\sum_{j=1}^{n}\nu_{\theta}(X_j)R_{n,3}(X_j) + O(h\delta_n)\\ &= \frac{1}{n}\sum_{j=1}^{n}\nu_{\theta}(X_j)\{\frac{1}{n}f_{\theta}^{-1}(\theta^T X_j)\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)\varepsilon_i\} + O(h\delta_n)\\ &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{n}\sum_{j=1}^{n}K_{h,i}^{\theta}(X_j)\nu_{\theta}(X_i)f_{\theta}^{-1}(\theta^T X_j)\Big\}\varepsilon_i + O(h\delta_n)\\ &= \frac{1}{n}\sum_{i=1}^{n}\nu_{\theta}(X_i)\varepsilon_i + \frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{n}\sum_{j=1}^{n}K_{h,i}^{\theta}(X_j)\nu_{\theta}(X_i)f_{\theta}^{-1}(\theta^T X_j) - \nu_{\theta}(X_j)\Big\}\varepsilon_i + O(h\delta_n)\\ &= \frac{1}{n}\sum_{i=1}^{n}\nu_{\theta_0}(X_j)\varepsilon_i + O(h\delta_n). \qquad (6.42)\end{aligned}$$

Similarly,

$$\begin{aligned}&\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)Z_iX_{ij}^T\theta_0R_{n,4}(X_j)/\varsigma_0^{\theta}(\theta_0^T X_j)\\ &= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)Z_iX_{ij}^T\theta_0f_{\theta}^{-2}(\theta^T X_j)\frac{1}{nh^2}\sum_{\ell=1}^{n}K_{h,\ell}^{\theta}(X_j)\theta^T X_{\ell j}\varepsilon_{\ell}\\ &\qquad + O\{(h^2+\delta_{\theta})h^{-1}\delta_n\tau_n\}\\ &= O(\delta_n^2 + h\delta_n + h^{-1}\delta_n\delta_{\theta}). \qquad (6.43)\end{aligned}$$

Combining (6.41)-(6.43) and Lemma A.2, we have

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)Z_i\tilde{\Delta}_{1,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = \frac{1}{n}\sum_{i=1}^{n}\{Z_i - \nu_{\theta_0}(X_j)\}\varepsilon_i + O(h\delta_n + h^{-1}\delta_n\delta_{\theta}).$$

By (C2), μ_{θ} has bounded derivative in θ. Hence

$$\frac{1}{n}\sum_{i=1}^{n}K_{h,i}^{\theta}(x)Z_i\tilde{\Delta}_{3,i}^{\theta} = f_{\theta_0}(\theta_0^T x)g'(\theta_0^T x)\nu_{\theta_0}\{\mu_{\theta_0} - x\}^T\theta_d + O\{(J_{\theta}+\delta_{\theta}+\tau_n)\delta_{\theta}\},$$

Since $E[g'(\theta_0^T X)\nu_{\theta_0}(X)\{\mu_{\theta_0}(X) - X\}] = 0$, we have

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) Z_i \tilde{\Delta}_{3,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = O\{(\delta_\theta + h + \delta_n)\delta_\theta\}.$$

It is easy to see that

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} Z_i \tilde{\Delta}_{4,i}^{\theta} = \nu_\theta \nu_\theta^T \beta_d + O\{(J_\theta + \delta_n)\delta_\beta\}.$$

By the first part of Lemma A.1, we have

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) Z_i \tilde{\Delta}_{4,i}^{\theta}(X_j) = E\{\nu_{\theta_0}(X)\nu_{\theta_0}^T(X)\}\beta_d + O\{(h + \delta_n)\delta_\beta\}.$$

For the other terms, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} Z_i \tilde{\Delta}_{2,i}^{\theta} = O(\delta_\theta^2 + h\delta_\theta + h^3 + h^2 J_\theta),\\
&\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} Z_i \tilde{\Delta}_{5,i}^{\theta} = O\{(h^2 + \delta_\theta^2)(h + J_\theta) + \delta_n(\delta_\theta^2 + h^2)\},\\
&\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} Z_i \tilde{\Delta}_{6,i}^{\theta} = O(\delta_\theta^2 + h^{-1} + hJ_\theta\delta_\gamma + \delta_n\delta_\gamma + h\tau_n),\\
&\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} Z_i \tilde{\Delta}_{7,i}^{\theta} = O\{(\delta_\theta + h + h^{-1}\delta_n)\delta_\gamma + h\tau_n\}.
\end{aligned}$$

By (6.24), we have

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) Z_i \{\tilde{\Delta}_{2,i}^{\theta}(X_j) + \tilde{\Delta}_{5,i}^{\theta}(X_j) + \tilde{\Delta}_{6,i}^{\theta}(X_j) + \tilde{\Delta}_{7,i}^{\theta}(X_j)\} = O\{(h + \delta_n)\delta_\beta\}.$$

Combining the forgoing equations, we finish the proof of the third part of Lemma 6.

Note that $R_{n,4} = O(h^{-1}\delta_n)$. Hence

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} R_{n,4} X_{i0}^T \theta_0 = O\{h^{-1}\delta_n\delta_\theta + (h + J_\theta)\delta_n\}.$$

By (6.24), we have

$$\begin{aligned}
\frac{1}{n^2}\sum_{j=1}^{n}\tilde{d}_j\sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij} \tilde{\Delta}_{\theta,1,i}/\varsigma_0^{\theta}(\theta^T X_j) &= \frac{1}{n^2}\sum_{j=1}^{n} g'(\theta_0^T X_j) f_\theta^{-1}(\theta^T X_j)\\
&\times \sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij}\Big\{\varepsilon_i - \frac{1}{n}\sum_{\ell=1}^{n} K_{h,l}^{\theta}(X_j)\varepsilon_\ell\Big\} + O\{(h + h^{-1}\delta_n)\delta_\theta + h\delta_n\}.
\end{aligned}$$

By the set of equations in (6.32), we have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{j=1}^{n} g'(\theta_0^T X_j) f_\theta^{-1}(\theta^T X_j) \sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij} \{\frac{1}{n}\sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(X_j)\varepsilon_\ell\} \\
&= \frac{1}{n}\sum_{j=1}^{n} g'(\theta_0^T X_j)\{\mu_\theta(X_j) - X_j\}\{\frac{1}{n}\sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(X_j)\varepsilon_\ell\} + O\{(h+\delta_n)\delta_n\} \\
&= \frac{1}{n}\sum_{j=1}^{n} g'(\theta^T X_j)\{\mu_\theta(X_j) - X_j\}\{\frac{1}{n}\sum_{\ell=1}^{n} K_{h,\ell}^{\theta}(X_j)\varepsilon_\ell\} + O\{(h+\delta_n+\delta_\theta)\delta_n\} \\
&= O(h\delta_n + \delta_n\delta_\theta).
\end{aligned}$$

Therefore, by Lemma A.2 and the third equation of Lemma A.1, we have

$$\begin{aligned}
&\frac{1}{n^2}\sum_{j=1}^{n} \tilde{d}_j \sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij} \tilde{\Delta}_{1,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) \\
&= \frac{1}{n^2}\sum_{j=1}^{n} g'(\theta_0^T X_j) f_\theta^{-1}(\theta^T X_j) \sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij}\varepsilon_i + O(h^{-1}\delta_n\delta_\theta + h\delta_n + h^{-1}\delta_n^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} \{\frac{1}{n}\sum_{j=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij} g'(\theta^T X_j) f_\theta^{-1}(\theta^T X_j)\}\varepsilon_i + O(h^{-1}\delta_n\delta_\theta + h\delta_n + h^{-1}\delta_n^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} g'(\theta^T X_i)\{\mu_\theta(X_i) - X_i\}\varepsilon_i + O(h^{-1}\delta_n\delta_\theta + h\delta_n + h^{-1}\delta_n^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} g'(\theta_0^T X_i)\{\mu_{\theta_0}(X_i) - X_i\}\varepsilon_i + O(h^{-1}\delta_n\delta_\theta + h\delta_n + h^{-1}\delta_n^2). \qquad (6.44)
\end{aligned}$$

By the equations in (6.32), we have

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} X_{i0} \tilde{\Delta}_{3,i}^{\theta} = -f_\theta g'(\theta_0^T x)\{\mu_\theta - x\}\{\mu_\theta - x\}^T\theta_d + O\{\delta_\theta(h + J_\theta + \delta_n)\}.$$

Note that $\tilde{d}_j = g'(\theta_0^T X_j) + O\{(1 + h^{-1}J_\theta)\delta_\gamma + h^{-1}\delta_n\}$. We have by (6.24),

$$\begin{aligned}
&\frac{1}{n}\sum_{j=1}^{n} \tilde{d}_j \sum_{i=1}^{n} K_{h,i}^{\theta}(X_j) X_{ij} \tilde{\Delta}_{3,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) \\
&= \frac{1}{n}\sum_{j=1}^{n} \{g'(\theta_0^T X_j)\}^2\{\mu_\theta(X_j) - X_j\}\{\mu_\theta(X_j) - X_j\}^T\theta_d + O\{\delta_\theta(\delta_\gamma + h^{-1}\delta_n) + \delta_\theta\tau_n\} \\
&= \frac{1}{n}\sum_{j=1}^{n} \{g'(\theta_0^T X_j)\}^2\{\mu_{\theta_0}(X_j) - X_j\}\{\mu_{\theta_0}(X_j) - X_j\}^T\theta_d + O\{\delta_\theta(\delta_\gamma + h^{-1}\delta_n) + \delta_\theta\tau_n\} \\
&= \tilde{W}_0\theta_d + O\{\delta_\theta(\delta_\gamma + h^{-1}\delta_n) + \delta_\theta\tau_n\}. \qquad (6.45)
\end{aligned}$$

The first part of Lemma A.1 was used to obtain the last equation above. Similarly,

$$\frac{1}{n}\sum_{i=1}^{n} K_{h,i}^{\theta} X_{i0} \tilde{\Delta}_{4,i}^{\theta} = f_\theta\{\mu_\theta - x\}\nu_\theta^T\beta_d + O\{(J_\theta + \delta_n)\delta_\beta\}.$$

Note that $E\{\mu_\theta(X_j) - X_j\}\nu_\theta^T(X_j) = 0$. By the first part of Lemma A.1, we have

$$\frac{1}{n}\sum_{j=1}^{n}\{\mu_\theta(X_j) - X_j\}\nu_\theta^T(X_j) = O(\delta_{0n}).$$

By (6.24), we have

$$\frac{1}{n}\sum_{j=1}^{n}\tilde{d}_j\sum_{i=1}^{n}K_{h,i}^{\theta}(X_j)X_{ij}\tilde{\Delta}_{4,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = O\{(\delta_\gamma + h + h^{-1}\delta_n)\delta_\beta\}. \tag{6.46}$$

For the other terms, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}K_{h,i}^{\theta}X_{i0}\tilde{\Delta}_{2,i}^{\theta} &= \frac{1}{2}g''(\theta_0^T x)\Big\{S_{1,2}^{\theta} + 2\theta_d^T S_{2,1}^{\theta} + S_3^{\theta} - S_1^{\theta}h^2\Big\} \\
&= O(h^2\delta_\theta + h^2\tau_n + h^2 J_\theta + \delta_\theta^2), \\
\frac{1}{n}\sum_{i=1}^{n}K_{h,i}^{\theta}X_{i0}\tilde{\Delta}_{5,i}^{\theta} &= O(\delta_\theta^2 + hJ_\theta\delta_\theta + h^2\delta_\theta), \\
\frac{1}{n}\sum_{i=1}^{n}K_{h,i}^{\theta}X_{i0}\{\tilde{\Delta}_{6,i}^{\theta} + \tilde{\Delta}_{7,i}^{\theta}\} &= O\{\delta_\theta^2 + (h^{-1}\tau_n + J_\theta)\delta_\gamma + h\tau_n\}.
\end{aligned}$$

Thus

$$\frac{1}{n}\sum_{j=1}^{n}\tilde{d}_j\sum_{i=1}^{n}K_{h,i}^{\theta}X_{i0}\tilde{\Delta}_{2,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = O(h\delta_\theta + h^2\tau_n + \delta_\theta^2), \tag{6.47}$$

$$\frac{1}{n}\sum_{j=1}^{n}\tilde{d}_j\sum_{i=1}^{n}K_{h,i}^{\theta}X_{i0}\tilde{\Delta}_{k,i}^{\theta}(X_j)/\varsigma_0^{\theta}(\theta^T X_j) = O\{(\delta_\gamma + h + h^{-1}\tau_n)\delta_\gamma + h\tau_n\}, \tag{6.48}$$

$k = 5, 6, 7$. Therefore the last part of Lemma A.6 follows from (6.44)-(6.48). □

Proof of Lemma 1. We shall prove that the equations in the Lemma 1 hold with probability 1. Therefore, Lemma A.1 follows. From Lammas A.5 and A.6 and (6.29), we have for any β and θ with $\theta^T\theta = 1$,

$$\bar{\beta} - \beta_0 = \{E(ZZ^T)\}^{-1}E\{\nu(Z)\nu^T(Z)\}(\beta - \beta_0) + O(b + b^{-1}\delta_{pn}). \tag{6.49}$$

Note that the above equation does not depend on the choice of θ. This is because we use a multivariate kernel, i.e. we use a more general multivariate function to replace $g(\theta_0^T x)$. In the algorithm, (6.49) can be written as

$$\bar{\beta}_{k+1} - \beta_0 = \{E(ZZ^T)\}^{-1}E(\nu(X)\nu^T(X))(\bar{\beta}_k - \beta_0) + O(b + b^{-1}\delta_{pn}), \tag{6.50}$$

where the sub-index k indicates that the corresponding values are the results of the k'th iteration in the algorithm; see (6.25) and (6.26). By assumption (C6), $E(ZZ^T) - E\{\nu(X)\nu^T(X)\}$

is a positive definite matrix. Note that $E\{\nu(X)\nu^T(X)\}$ is a semipositive matrix. Hence the eigenvalues of $\{E(ZZ^T)\}^{-1}E\{\nu(X)\nu^T(X)\}$ are all less than 1. After sufficiently many steps, we have from (6.50)

$$\bar{\beta}_k - \beta_0 = O(b + b^{-1}\delta_{pn}). \tag{6.51}$$

See the proof of Theorem 1 below for more details. Therefore

$$\tilde{\beta} - \beta_0 = O(b + b^{-1}\delta_{pn}). \tag{6.52}$$

If $\theta^T\theta_0 \neq 0$, then it follows from Lemmas A.5 and A.6 and (6.29) that

$$\bar{\theta} - \theta_0 = (\theta^T\theta_0)^{-1}(1 - \theta^T\theta_0)\theta_0 + O(\delta_\beta + b + b^{-1}\delta_{pn}),$$

i.e. $\bar{\theta} = (\theta^T\theta_0)^{-1}\theta_0 + O(\delta_\beta + b + b^{-1}\delta_{pn})$. By (6.52), we may assume δ_β is small enough (otherwise, take $\beta = \tilde{\beta}$). Thus

$$\bar{\theta} =: \bar{\theta}/|\bar{\theta}| = sign(\theta^T\theta_0)\theta_0 + O(\delta_\beta + b + b^{-1}\delta_{pn}).$$

In the algorithm, we have

$$\bar{\theta}_{k+1} - sign(\theta^T\theta_0)\theta_0 = O(\delta_{\bar{\beta}_k} + b + b^{-1}\delta_{pn}). \tag{6.53}$$

Combining (6.51) and (6.53), we have,

$$\tilde{\theta} - sign(\theta^T\theta_0)\theta_0 = O(b + b^{-1}\delta_{pn}). \tag{6.54}$$

The proof is completed. □

Proof of Theorem 1. We only prove the first part in the case $\tilde{\theta}^T\theta_0 > 0$. The second part follows immediately from the first part and Theorem 1 of Carroll *et al.* (1997). It follows from Lemmas A.1, A.4 and A.5 and (6.30) that

$$\begin{pmatrix} \tilde{\beta} - \beta_0 \\ \tilde{\theta} - \theta_0 \end{pmatrix} = \tilde{D}^- N_n + \tilde{D}^-\tilde{C}\begin{pmatrix} \beta - \beta_0 \\ \theta - \theta_0 \end{pmatrix} + O\{(\delta_\gamma + h + h^{-1}\delta_n)\delta_\gamma + h\tau_n + h^{-1}\delta_n^2\}, \tag{6.55}$$

where

$$\tilde{C} = \begin{pmatrix} E\{\nu_{\theta_0}(X)\nu_{\theta_0}^T(X)\} & 0 \\ 0 & \tilde{W}_0 \end{pmatrix}, \qquad \tilde{D} = \begin{pmatrix} E(ZZ^T) & \tilde{C}_{12} \\ \tilde{C}_{12}^T & 2\tilde{W}_0 \end{pmatrix},$$
$$N_n = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} Z_i - \nu(X_i|\theta_0) \\ g'(\theta_0^T X_i)\{\mu_{\theta_0}(X_j) - X_i\} \end{pmatrix}\varepsilon_i.$$

Following the proof of Lemma 1 of Xia *et al.* (1999), we have $\tilde{C}$, $\tilde{D}$ and $W_0 = \tilde{D} - \tilde{C}$ are all semi-positive matrices with rank $p+q-1$. Therefore, $D \stackrel{def}{=} (\tilde{D}^-)^{1/2}\tilde{C}(\tilde{D}^-)^{1/2}$ is a semi-positive matrix with all eigenvalues less than 1. There exist $1 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p+q-1} > 0$ and orthogonal matrix Γ such that

$$D = \Gamma diag(\lambda_1, \cdots, \lambda_{p+q-1}, 0)\Gamma^T.$$

Let $(\tilde{\beta}_k, \tilde{\theta}_k)$ be the calculation results of the k'th iteration in the algorithm; see (6.27) and (6.28). For any k, equation (6.55) holds with $(\tilde{\beta}, \tilde{\theta})$ replaced by $(\tilde{\beta}_{k+1}, \tilde{\theta}_{k+1})$ and (β, θ) by $(\tilde{\beta}_k, \tilde{\theta}_k)$. Let $\tilde{\gamma}_k = \tilde{D}^{1/2}(\tilde{\beta}_k^T - \beta_0^T, \tilde{\theta}_k^T - \theta_0^T)^T$, we have

$$\tilde{\gamma}_{k+1} = (\tilde{D}^-)^{1/2}N_n + D\tilde{\gamma}_k + O\{(\delta_{\tilde{\gamma}_k} + h + h^{-1}\delta_n)\delta_{\tilde{\gamma}_k} + h\tau_n + h^{-1}\delta_n^2\}. \tag{6.56}$$

It follows that

$$\begin{aligned} \delta_{\tilde{\gamma}_{k+1}} &\leq \delta_{0n} + \lambda_1\delta_{\tilde{\gamma}_k} + c(\delta_{\tilde{\gamma}_k} + h + h^{-1}\delta_n)\delta_{\tilde{\gamma}_k} + c(h\tau_n + h^{-1}\delta_n^2) \\ &= \delta_{0n} + \{\lambda_1 + c\delta_{\tilde{\gamma}_k} + c(h + h^{-1}\delta_n)\}\delta_{\tilde{\gamma}_k} + c(h\tau_n + h^{-1}\delta_n^2) \end{aligned} \tag{6.57}$$

almost surely, where c is a constant. We can further take $c > 1$. Because $h, h^{-1}\delta_n, \tau_n, \delta_{0n} \to 0$ as $n \to \infty$, we may assume that

$$c(h + h^{-1}\delta_n) \leq (1-\lambda_1)/3, \quad \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2) \leq (1-\lambda_1)^2/(9c). \tag{6.58}$$

By (6.52) and (6.54), we may assume

$$\delta_{\tilde{\gamma}_1} \leq (1-\lambda_1)/(3c). \tag{6.59}$$

Therefore, it follows from (6.57), (6.58) and (6.59) that

$$\delta_{\tilde{\gamma}_2} \leq \{\lambda_1 + 2(1-\lambda_1)/3\}(1-\lambda_1)/(3c) + (1-\lambda_1)^2/(9c) = (1-\lambda_1)/(3c). \tag{6.60}$$

From (6.57), (6.58) and (6.60), we have

$$\delta_{\tilde{\gamma}_3} \leq (1-\lambda_1)/(3c).$$

Consequently, $\delta_{\tilde{\gamma}_k} \leq (1-\lambda_1)/(3c)$ for all k. Therefore we have from (6.57) that

$$\delta_{\tilde{\gamma}_{k+1}} \leq \lambda_0\delta_{\tilde{\gamma}_k} + \delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)$$

almost surely, where $0 \leq \lambda_0 < (2+\lambda_1)/3 < 1$. It follows that

$$\delta_{\tilde{\gamma}_k} \leq \lambda_0^k\delta_{\tilde{\gamma}_1} + \{\delta_{0n} + c(h\tau_n + h^{-1}\delta_n^2)\}\sum_{j=1}^{\infty}\lambda_0^j = O(\delta_{0n} + h\tau_n + h^{-1}\delta_n^2),$$

for sufficiently large k. By (6.56), we have

$$\begin{aligned}\tilde{D}^{1/2}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\theta}-\theta_0\end{pmatrix} &= (\tilde{D}^{-})^{1/2}N_n + D\tilde{D}^{1/2}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\theta}-\theta_0\end{pmatrix} + O(\delta_{0n}^2 + h\tau_n + h^{-1}\delta_n^2)\\ &= (\tilde{D}^{-})^{1/2}N_n + D\tilde{D}^{1/2}\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\theta}-\theta_0\end{pmatrix} + o(n^{-1/2}). \qquad (6.61)\end{aligned}$$

The facts that $n^{1/2}h^3 \to 0$ and $n^{1/2}h^{-1}\delta_n^2 \to 0$ are used in the last step above. It follows from (6.61) that

$$(\tilde{D} - \tilde{D}^{1/2}D\tilde{D}^{1/2})\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\theta}-\theta_0\end{pmatrix} = N_n + o(n^{-1/2}),$$

or

$$W_0\begin{pmatrix}\hat{\beta}-\beta_0\\ \hat{\theta}-\theta_0\end{pmatrix} = N_n + o(n^{-1/2}).$$

The first part of Theorem 1 follows from the above equation and the central limiting theorem of dependent data, see e.g. Rio (1995). □

ACKNOWLEDGEMENTS

The first author thanks the Friends of London School of Economics (Hong Kong) and the Wellcome Trust for partial support. The research has been partially supported by Sonderforschungsbereich 373, Berlin.

REFERENCES

Arminger, G., Euache, D. and Bonne, T (1997). Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feed forward subroles. *Comp. Statistics*, 12, 293 - 310.

Bradley, R.C.(1986) Basic Properties of strong mixing conditions. *In Dependence in probability and Statistics: A survey of Resent Results*, Ed. E. Eberlein and M.S.Taqqu, pp. 165-92. Boston: Birhauser.

Carroll, R.J., Fan. J. Gijbels, I. and Wand, M.P. (1997) Generalized partially linear single-index models. *J. Am. Statist. Ass.,* **92**, 477-489.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications.* Chapman & Hall, London.

Hall, P. (1989). On projection pursuit regression. *Ann. Statist.*, **17**, 573-588.

Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157-178.

Härdle, W., Janssen, P. and Serfling, R. (1988) Strong uniform consistency rates for estimators of conditional functionals. *Ann. Statist.* **16**, 1428-1449.

Härdle, W. and Stoker, T. M. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Amer. Stat. Ass.* **84** 986-995.

Henley, W. E. and Hand, D. J. (1996). A k-nearest neighbour classifier for assessing consumer credit risk, 45, 77 - 95.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001a) Direct estimation of the single-index coefficients in single-index models. *Ann. Statist.*, **29**, 1537 - 1566.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001b) Structure adaptive approach for dimension reduction. *Ann. Statist.* (to appear).

Ichimura, H. and Lee, L. (1991) Semiparametric least squares estimation of multiple index models: Single equation estimation. *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, edited by Barnett, W., Powell, J. and Tauchen, G.. Cambridge University Press.

Li, K. C. (1991) Sliced inverse regression for dimension reduction (with discussion). *Amer. Statist. Ass.*, **86**, 316-342.

Linton, O. (1995) Second order approximation in the partially linear regression model. *Econometrica*, **63**, 1079-1112.

Müller, M. and Rönz, B. (2000). Credit Scoring using semiparametric methods. In "Measuring Risk in Complex Stochastic Systems", Franke, Härdle, Stahl (eds.), Springer Lecture Notes in Statistics 147, p. 85 - 102.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications.* John Wiely & Sons.

Rio, E. (1995) The functional law of the iterated logarithm for stationary strongly mixing sequences. *Ann. Prob.*, **23**, 1188-1203.

Robinson, P. M. (1988) Root-N-Consistent semiparametric regression, *Econometrica*, **56**, 931-954.

Ruppert, D., Sheather, J., and Wand, P. M. (1995) An effective bandwidth selector for local least squares regression. *J. Am. Statist. Ass.*, **90**, 1257-1270.

Samarov, A., Spokoiny, V. and Vial, C. (2002) Component identification and estimation in nonlinear high-dimensional regression models by structure adaptation. Manuscript. Weierstrass Institute and Humboldt University.

Severini, T. A. and Staniswalis, I. G. (1994). Quasi-likelihood estimation in semiparametric models. Journal American Statistical Association 89, 501 - 511.

Xia, Y. and Li, W. K. (1999) On single-index coefficient regression models. *J. Amer. Statist. Ass.* **94**, 1275-1285.

Xia, Y., Tong, H. and W. K. Li (1999) On extended partially linear single-index models. *Biometrika*, 86, 831-842.

Xia, Y., Tong, H., Li, W. K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space (with discussions). *J. R. Statist. Soc. B.*, **64**, 1-28.